

DOCUMENT RESUME

ED 192 799

IR 008 881

AUTHOR Caplan, Priscilla Louise
TITLE Thesaurus-Based Automatic Indexing: A Study of Indexing Failure.
PUB DATE Dec 78
NOTE 129p.; Master's Thesis, University of North Carolina, Chapel Hill.
EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Automatic Indexing; Coordinate Indexes; *Failure; Information Retrieval; *Systems Analysis; *Thesauri
IDENTIFIERS Natural Language

ABSTRACT

This study examines automatic indexing performed with a manually constructed thesaurus on a document collection of titles and abstracts of library science master's papers. Errors are identified when the meaning of a posted descriptor, as identified by context in the thesaurus, does not match that of the passage of text which occasioned the posting. Three major types of errors are discussed: term ambiguity, flaws in thesaurus construction, and the automatic word matching procedure. The study concludes with a discussion of reasons for failure of each of the three error types, and an explanation of study limitations. Appendices contain an outline of the hierarchical structure of the thesaurus studied, a random sample of 100 index terms, and other subsets of the thesaurus which support the study. (Author/SW)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED192799

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

THESAURUS-BASED AUTOMATIC INDEXING:

A STUDY OF INDEXING FAILURE

By

Priscilla Louise Caplan

A master's paper submitted to the faculty of the School of Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science

Chapel Hill

December, 1978

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Priscilla L. Caplan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Approved by:

Merton Siller
Adviser

IR008881

PRISCILLA LOUISE CAPLAN. Thesaurus-based Automatic Indexing: A Study of Indexing Failure. A master's paper for the M.S. in L.S. degree. December, 1978. 121 pages. Advisor: M. Dillon.

This study examines automatic indexing performed with a manually constructed thesaurus on a document collection of titles and abstracts of library science master's papers. Errors are identified when the meaning of a posted descriptor, as identified by context in the thesaurus, does not match that of the passage of text which occasioned the posting. Three major reasons for failure are discussed: term ambiguity, flaws in thesaurus construction, and the automatic word matching procedure.

Headings:

Indexing -- Evaluation

Indexing -- Systems

Mechanical indexing

Thesauri

CONTENTS

PART ONE

I. THESAURUS-BASED AUTOMATIC INDEXING	1
Indexing errors of commission	3
The indexing: background information	5
II. METHOD	7
Identification of errors	7
The classification of errors	9
Comparison of error-causing terms with random sample of thesaurus terms	12

PART TWO

I. ERRORS RELATED TO TERM AMBIGUITY	17
Classification of ambiguous terms	18
Disambiguation of homographs	30
II. ERRORS RELATED TO THESAURUS CONSTRUCTION	60
Errors of false synonymy	60
Errors relating to the precision of indexing	67
Miscellaneous errors	71
III. ERRORS RELATED TO THE MATCHING PROCESS	76
False term combinations	76
Input errors	78
IV. CONCLUSION	79

APPENDICES

A. OUTLINE OF THE HIERARCHICAL STRUCTURE OF THE "MASTER'S PAPERS THESAURUS"	84
B. PHRASE MATCH CRITERIA	85

C.	RANDOM SAMPLE OF 100 INDEX TERMS IN THE "MASTER'S PAPERS THESAURUS"	87
D.	CONTEXT OF INDEX TERMS INVOLVED IN TYPE-A ERRORS	92
E.	HAZARDOUS TERMS IN THE "MASTER'S PAPERS THESAURUS"	102
F.	PRE- AND POST-COORDINATION OF TERMS	107
G.	SEMANTIC TERM RELATIONS WITHIN CONCEPT CLASSES	114

TABLES

1. Index Terms Involved in Errors and Distribution of Errors among Categories	13
2. Distribution of Types of Index Terms in Random Sample and in "Master's Papers Thesaurus"	16
3. Errors Related to True and to Contextual Ambiguity	25
4. Errors Related to True and to Contextual Ambiguity-- Literature* Omitted	26
5. Commission of Type-A Errors--Technical and Non-technical Terms	28
6. Singular and Plural Word Forms--Actual and Possible Use	36
7. Syntactic Tagging--Error-causing Terms	45
8. Precoordination and Term Matches--Random Sample Index Terms	47
9. Multiple-posting of Candidate Terms	55
10. Determination of Errors Involving the Descriptor <u>Children</u>	70
11. Errors Relating to Precision of Indexing	72
12. Distribution of Indexing Errors among Major Categories and Sub-categories	81
13. Terms Involved in Type-A Errors--Context of All Usages Occasioning No Errors or Type-A Errors	93
14. Classification of Hazardous Terms into Formal Homographs and Polysemes	104
15. Classification of Hazardous Terms into Truly and Contextually Ambiguous Terms	105
16. Precoordinated Index Terms which Found Matches in the Document Collection and Co-occurrence of Uniterm Components	109
17. Precoordinated Index Terms in Random Sample Finding No Matches in Document Collection and Co-occurrence of Uniterm Components	111
18. Synonym-Descriptor Relations--Error-causing Terms	115
19. Synonym-Descriptor Relations--Random Sample	117

PART ONE

I. THESAURUS-BASED AUTOMATIC INDEXING

This paper presents a study of indexing by a thesaurus-based automatic indexing system. Automatic indexing has been defined (Stevens 1965, p. 3) as "the use of machines to extract or assign index terms without human intervention once programs or procedural rules have been established." In thesaurus-based automatic indexing, portions of natural-language text are compared with a mechanized authority list. When a match is found between a thesaurus term and the text, a subject heading is automatically assigned to the document, or "posted." The thesaurus, then, must contain all of the words and phrases for which matches are sought in the text of the documents. In this paper, these words and phrases are referred to as "index terms."

Not all index terms will actually be assigned as subject headings. Like thesauri intended for manual indexing, the mechanized thesaurus identifies synonyms, spelling variations, and other terms considered equivalent for indexing purposes, and designates a single preferred term to represent them all. This index term is referred to here as the "descriptor," or "main term," and is posted when a match is found. The set of all index terms represented by one descriptor is known as a concept class.

Semantic relationships between concept classes are also noted in the thesaurus. Descriptors are arranged hierarchically, with more general terms hierarchically superior and more specific terms, inferior.

This information may be exploited in retrieval, to facilitate generic searches.

The thesaurus is a retrieval aid as well as an indexing tool, so it must have a printed as well as a machine-readable version. The thesaurus employed in this study was printed in two parts: a hierarchical display of descriptors, and an alphabetical index to the terms in the hierarchy. The alphabetical list indicates the grouping of index terms into concept classes, and makes cross-references between related descriptors. It also contains "See" references from index terms not used as descriptors to the appropriate main term. These index terms are known as the "lead-in" vocabulary, as they lead the user to the descriptors which actually point to documents.

Unlike thesauri used for manual indexing, mechanized thesauri contain some index terms which do not appear in the printed thesaurus because they are neither descriptors nor useful lead-in terms. For example, a user looking up the word 'catalogue' can easily find the descriptor Catalog in the alphabetical index without the aid of a "See" reference. The computer, however, needs to be informed that the two words are synonymous. 'Catalogue', then, appears in the machine-readable version for matching purposes only. These "non-print" terms constitute the main difference between the two versions of the thesaurus. A sample thesaurus record containing non-print terms is shown in figure 1.

Thesauri perform three basic functions. They determine which concepts in a document may be accessible through indexing by designating a limited number of concept classes. They help to bring the language of the searcher and the language of the documents into coincidence by providing control of synonyms and variations in term usage.

THESAURUS FILE

DOCUMENT FILE

Record 1.

DESCRIPTOR = Indexing

Record 2.

DESCRIPTOR = Automatic indexing
 LEAD-IN = Mechanical indexing
 LEAD-IN = Computerized indexing
 NON-PRINT = Mechanized indexing
 NON-PRINT = Automated indexing

Document 1.

. . . This study of a computer-
 ized thesaurus-based indexing
 system demonstrates one method
 for the evaluation of thesauri
 designed for mechanized
indexing. . . .

Descriptors posted

1. Indexing
2. Automatic indexing

Fig. 1. Index terms in machine-readable thesaurus and sample of indexed text. Terms which matched are underlined.

Finally, they guide users in the formulation of search strategies by indicating semantic relationships between descriptors.

Indexing Errors of Commission

The overall performance of an indexing system is a product of many complex and interrelated factors. These include the extent to which the thesaurus meets the conceptual requirements of users, the effectiveness of the index terms in distinguishing relevant from non-relevant material, and even the format by which the indexing language is displayed. We are concerned here with only one aspect of the indexing process: the coincidence of the natural language of the document collection and the controlled vocabulary of the thesaurus designed for computerized use.

Indexing failures arise when the meaning of a descriptor assigned to a document is not the same as the meaning of the portion of text which occasioned the posting of that descriptor. These instances are designated "errors of commission" here, for the text is

4

indexed, but inappropriately. A second kind of failure, or an "error of omission," occurs when a concept represented in both the thesaurus and the text is not indexed at all. Only errors of commission were considered in this investigation. The research was designed to indicate how often these errors occur, what causes them to occur, and how they might be prevented.

This line of investigation assumes that there is a particular meaning associated with any word or phrase in a thesaurus or a text, and that this meaning can be determined with a fair degree of precision and consensus. For most words appearing in context in a document, correct and incorrect interpretations can generally be agreed upon, in accordance with what is understood to be the intended meaning of the author. We will refer to what would be accepted as the intended interpretation as the "document-contextual" meaning of the term.

For terms which appear in a thesaurus, the meaning intended by the "author," or thesaurus design team, is a less useful standard than the meaning expected by the user. In this study, the meaning of any thesaurus term is taken to be a function of the reasonable expectations of a hypothetical user group. These expectations can be influenced by the thesaurus itself in several ways apart from the explicit provision of term definitions. If the thesaurus as a whole deals with a limited subject area, the import of many otherwise ambiguous words may be assumed. The expected meaning of the homograph 'pitcher', for example, in a thesaurus dealing with handicrafts and minor arts, is not likely to vary from user to user. Expectations are also affected by the position of a term in a hierarchy. If 'pitcher' appears as a descendant of 'baseball', the user will expect the term to apply to Tom

Seaver but not to wide-mouthed vases, regardless of the subject area of the thesaurus. Finally, expectations are influenced by the information provided about equivalent and related terms: if 'pitcher' and 'beaker' are shown to belong to the same concept class, one expects the meanings to be fairly synonymous. Non-print index terms, of course, will never be seen by the user, and so will play no role in shaping his expectations. Lead-in vocabulary, on the other hand, can be very influential.

Although an actual user of the system could conceivably associate any meaning with any given term, the idea of "reasonable expectation" requires congruence with the logical implications of thesaurus organization. The organization provides a context to the index term which generally demarcates a subset of appropriate meanings from the set of all meanings associated with any natural language word or phrase. This range of acceptable meanings may be called the "thesaurus-defined" meaning of a term, as opposed to the "document-contextual" meaning of the same term as it appears in a document.

An indexing error of commission, then, occurs when a portion of text causes the posting of a descriptor, and the document-contextual meaning differs from the thesaurus-defined meaning of the word(s) involved.

The Indexing: Background Information

The indexing examined here was performed on the titles and abstracts of 149 master's papers written by candidates for the M.S. in L.S. degree at the School of Library Science at the University of North Carolina at Chapel Hill. Each title and abstract is referred to in the remainder of this paper as a "document"; the 149 title-abstract

set. stitute the "document collection." Indexing was performed using the Automatic Indexing system described in detail in Peggy S. Strode's "User's Manual for the Automatic Indexing System" (1976).

The library science "Master's Papers Thesaurus" was assembled by a team of five library science students in the spring of 1978, and an outline of its major hierarchical divisions appears in appendix A. For a more complete description of the thesaurus and technical information about the machine-readable format, see the "Thesaurus Maintenance Manual" (Silver 1974) and "Master's Papers Thesaurus" (Caplan et al. 1978).

Throughout this paper, words and phrases referred to as terms in the language or in document texts are designated in the usual manner, set off by single quotes. However, when a term included in the "Master's Papers Thesaurus" is mentioned in its specific capacity as an index term, the term is underlined and the initial letter capitalized (e.g. Catalog). When the term is referred to in both its singular and plural forms, truncation is indicated by an asterisk (e.g. Fine*, Librar*).

II. METHOD

In this study, the errors of commission in the thesaurus-based automatic indexing experiment were identified by comparing the thesaurus-defined and document-contextual meanings of all of the terms involved in matches. Errors were divided into three broad categories according to the underlying reasons for failure, and each class of errors was then analyzed in more detail.

Identification of Errors

In addition to the printed version of the thesaurus, four computer-produced reference tools were used to facilitate the identification of errors:

1. A printout of the master file for the machine-readable thesaurus, grouping all of the index terms in each concept class
2. An alphabetical listing of all descriptors which actually posted to one or more documents, giving for each descriptor the title of every document to which it posted
3. A printout of the full titles and abstracts of the 149 master's papers, and the descriptors assigned to each document
4. A Key Word in Context (KWIC) index of the titles and abstracts of the master's papers

All of the descriptors which had actually been assigned in the indexing were noted individually, along with the documents to which they were assigned and any other index terms in the same concept class. The

documents indexed by each descriptor were then scanned for the occurrence of any of the index terms in the concept class of the descriptor involved. This was necessary because one can not tell from the posted descriptor itself whether that term or some other member of the same concept class actually matched a term in the text.

Several features of the Automatic Indexing system as it was used for the master's papers collection complicate the identification of term matches. Before any matches are sought, the text is normalized by converting all letters to upper case, stripping punctuation and deleting numbers. In addition, plural words are converted to singular forms, and twenty stop words (common articles, prepositions, and the like) are deleted. Matched terms, therefore, are not always identical. The word form 'audio-visual', for example, will match with 'audio visuals', but not with 'audiovisual'.

In the matching procedure, longer terms are searched for in the text first, and if found, that portion of text is removed from consideration. The passage, 'library services to young adults', therefore, will be indexed by Services to young adults, precluding the posting of either Library services or Young adults.

Finally, the user may request different criteria for the matching of multiple-word phrases (see appendix B). The master's papers collection was indexed according to an option by which the words in a phrase may appear in any order but must be adjacent to each other in the text. Stop words, however, are disregarded entirely. The index term Library and society, then, will match 'library society', 'libraries of the society', 'society libraries', and other similar constructions.

Taking these features into consideration, all portions of text

which were found to match any index term in the concept class were compared semantically with the thesaurus-defined meaning of the descriptor. When the textual usage was included in the range of acceptable meanings represented by the descriptor, no further action was taken. When a significant discrepancy was found, the index term involved was noted along with enough of the document text to identify the passage responsible for the error.

The documents themselves were used in this process in order to ascertain more precisely the meanings of the passages under consideration. Once errors had been identified and categorized, however, the KWIC index was preferred as a more convenient representation of the vocabulary. All index terms involved in errors were looked up in the KWIC, and the total number of occurrences of each term in the document collection were counted and recorded, providing a check upon the error count, and an indication of the number of acceptable term matches.

The Classification of Errors

The criterion for determining indexing failures was the degree of overlap between the thesaurus-defined and document-contextual meanings of any words involved in term matches. When a number of errors had been isolated and examined, it was possible to identify three major factors underlying significant differences in meaning. These formed the basis for the initial categorization of errors.

The first category, designated as Type-A, consisted of errors related to the ambiguity of words in the language. For example, the index term Objective, with a thesaurus-defined meaning equivalent to 'aim' or 'goal', matched the same word used in a document as the antonym of 'subjective'. This was classed as a Type-A error.

Mistakes or omissions in the thesaurus itself were considered responsible for another group of errors, designated Type-B. The major source of these errors was the composition of concept classes. In some cases, a word or phrase which should have been included in a particular concept class was omitted, allowing an inappropriate descriptor to match the same word(s) in a document. For example, the phrase 'motion pictures' was indexed by the descriptor Pictures, because 'motion pictures' was not included as a synonym for Films.

A different kind of Type-B error occurred when concept classes contained inappropriate index terms. The descriptor Book reviewing, for example, posted to documents about record and film reviewing, because 'reviewer' was included in the concept class as a synonymous non-print index term.

Type-B errors were often difficult to identify, and some questionable cases were either reclassified or dismissed as the initial typology was clarified. No thesaurus can anticipate all of the ways any given concept may be expressed in the English language, and reasonable guidelines had to be established for determining when an omission constituted a thesaurus flaw. This problem will be discussed in detail in "Errors Related to the Precision of Indexing" below.

Different problems arose in the identification of Type-B errors involving inappropriate members of concept classes. Index terms are partially defined by the other term members of the same concept class, with all members contributing to the definition of the concept class as a whole. The difficulties this presents are illustrated by the example of Book reviewing. Does the presence of Reviewer in the same concept class effectively expand the meaning of the class as a

whole to encompass all sorts of reviews? Decisions were based upon the factors which would influence the reasonable expectations of the user: the normal semantic associations of the term, its position in the hierarchy, and, of course, whether or not the term was displayed as a descriptor or lead-in index term. As Reviewer was a non-print term incapable of influencing a user's expectations, its inclusion was considered an imperfection in the construction of the thesaurus. Other cases, however, were less clear cut.

A second problem arose in distinguishing errors involving thesaurus flaws from simple instances of term ambiguity. The word 'reviewer', for example, might pertain to book reviewing or not, depending upon its usage, and so could be categorized as an ambiguous term. The following test was applied: if the relocation of the index term in question to a different concept class or a different position in the hierarchy would serve to eliminate the errors under consideration, the thesaurus arrangement was considered faulty. Reviewer would have occasioned no errors as a hierarchical parent of Book reviewing, since all textual usages of the word did refer to reviews of various library materials. The word 'objective', on the other hand, was used with two different meanings, such that any thesaurus location appropriate to one would be inappropriate to the other. Errors occasioned by Objective were typed as Type-A, and those occasioned by Reviewer as Type-B.

The final class of errors, Type-C, was reserved for instances of indexing failure occasioned by the mechanical nature of the automatic indexing process itself. Mistakes in input precluded the matching of some words, while in other cases multiple-word descriptors matched with semantically unrelated word combinations. Book number, for example, a

descriptor synonymous with Cutter number, posted to a document which referred to "the number of books . . . written by women." Such phenomena involved neither term ambiguity nor flaws in the thesaurus itself, and accounted for only a small percentage of all errors identified.

The results of this classification are shown in table 1, which lists the 56 index terms involved in errors, and indicates the distribution of errors among them. These findings form the basis for the more detailed analysis of indexing failure presented in part two of this paper.

Comparison of Error-causing Terms With Random Sample of Thesaurus Terms

In the analysis of indexing errors, it was often useful to know the distribution of different types of terms, or of relationships between terms, in the thesaurus as a whole. For example, the total number of "technical" library science terms in the thesaurus had to be estimated in order to determine whether they were disproportionately represented among error-causing index terms (see "Technical and Non-technical Terms" below). To allow calculations of this nature, a random sample of 100 index terms (6% of all index terms) was taken. All index terms were numbered in order as they appeared in the thesaurus master file, and sample terms were drawn according to a table of random sampling numbers.

The terms included in the sample are listed and characterized individually in appendix C. Seven of the sample terms (7%) were involved in errors of commission identified in this study. Table 2 below compares the distribution of descriptors, lead-in terms and non-print terms in the sample with their distribution in the thesaurus as a whole.

TABLE 1.--Index terms involved in errors and distribution of errors among categories

Index term	Descriptor for concept class (if index term is not descriptor)	Errors			Matches
		Type-A	Type-B	Type-C	
Acquisition*	. . .	1	. . .	. . .	10
Administrative	Library director	. . .	2	. . .	2
Adult*	. . .	. . .	2	. . .	5
Aim*	Objectives	2	. . .	. . .	3
AMERICAN* LIBRAR*	. . .	1	. . .	1	2
Bibliograph*	. . .	2	. . .	1	21
Book* number*	. . .	. . .	. . .	1	1
Catalog*	. . .	4	3	. . .	17
Children	. . .	. . .	19	. . .	42
Classification*	. . .	4	. . .	. . .	6
Climate*	. . .	2	. . .	. . .	2
Contemporar*	. . .	1	. . .	. . .	11
Continuing education*	. . .	. . .	2	. . .	2
Diar*	Manuscripts	. . .	1	. . .	1
Dictionary*	. . .	1	. . .	. . .	1
Economic*	. . .	4	. . .	. . .	4
Education*	. . .	. . .	1	. . .	13
Environment*	Environmental science	. . .	4	. . .	5
Environmental	Environmental science	. . .	3	. . .	3
Fiction*	. . .	. . .	7	. . .	21

TABLE 1.--Continued

Index term	Descriptor for concept class (if index term is not descriptor)	Errors			Matches
		Type-A	Type-B	Type-C	
Fine*	Fine policies	1	. . .	. . .	4
Grant*	. . .	2	. . .	. . .	3
Index*	. . .	16	. . .	. . .	16
Information service*	Reference service	. . .	4	. . .	6
Institute*	Institutes and workshops	. . .	1	. . .	1
Interviewed by telephone*	Telephone reference	. . .	1	. . .	1
Journal*	Periodicals	1	. . .	. . .	11
Leadership	. . .	2	. . .	. . .	7
Level* of service*	Staffing	. . .	1	. . .	1
Librar* and societ*	. . .	. . .	. . .	1	1
Librar* building*	Library construction	2	. . .	. . .	3
Librar* law*	Library legislation	. . .	. . .	1	1
LIBRAR* LITERATURE*	. . .	4	. . .	. . .	5
Librar* science*	Library education	. . .	6	. . .	6
Literature*	. . .	57	2	. . .	64
Objective*	. . .	1	. . .	. . .	12
Periodical*	. . .	. . .	1	. . .	20
Picture*	. . .	4	3	. . .	10

TABLE 1--Continued

Index term	Descriptor for concept class (if index term is not descriptor)	Errors			Matches
		Type-A	Type-B	Type-C	
Philosoph*	. . .	1	. . .	. . .	1
Publication*	. . .	3	2	. . .	23
Reader*	Reading	4	. . .	. . .	12
Reading*	. . .	6	3	. . .	17
Record*	. . .	13	. . .	. . .	21
Religious	Churches	. . .	8	. . .	8
Reviewer*	Book reviewing	. . .	4	. . .	5
Science*	. . .	. . .	1	. . .	15
Selection*	. . .	3	5	1	27
Ser*(-ies)	. . .	2	. . .	. . .	8
State* librar*	. . .	. . .	1	. . .	13
Student*	. . .	2	. . .	. . .	29
Subject* analysis	. . .	1	. . .	. . .	1
Training method*	Staff training	. . .	1	. . .	1
User survey*	. . .	. . .	. . .	1	1
Wake Count* Public* Librar*	. . .	. . .	1	. . .	1
Western*	. . .	1	. . .	. . .	1
Young adult*	. . .	. . .	1	. . .	19
Total		133	90	7	547

TABLE 2.--Distribution of types of index terms in random sample and in
"Master's Papers Thesaurus"

	Sample	Thesaurus
Total terms	100	1650
Descriptors	47	753
Percentage of total	47%	45.6%
Lead-in terms	15	230
Percentage of total	15%	14.0%
Non-print terms	38	667
Percentage of total	38%	40.4%

PART TWO

I. ERRORS RELATED TO TERM AMBIGUITY

Written language can be ambiguous, or subject to more than one interpretation, in a variety of ways. A portion of text might reflect fuzzy thinking, or be organized in a confusing manner, or it could be unclear on a connotative level, as when the reader is uncertain whether irony is intended. Structural ambiguity, which most interests linguists, occurs at the sentence level, when a given combination of orthographic words represents more than one grammatical construction, as in the sentence, "Flying planes can be dangerous." This discussion is primarily concerned with term ambiguity, or the variety of meanings which can be associated with a single orthographic word or phrase.

Of 230 indexing errors identified in this study, 133 or 57.8%, were classified as errors involving term ambiguity. These errors were occasioned by thirty different index terms (see appendix D) which represent 53.6% of all terms involved in errors, and slightly less than 2% of all thesaurus index terms. These figures indicate that a relatively small number of homographs are responsible for a sizeable proportion of indexing errors.

A homograph may be informally defined as any word "with the same spelling as another but having a different meaning," (Aitchison and Gilchrist 1972, p. 18) and this is the usage that will be followed here unless otherwise noted. Manuals on thesaurus construction

generally advise design teams to omit homographs, or to treat them in special ways. When homographs are included, methods must be employed to reduce their bad effect on retrieval performance. According to Loll N. Rolling (1970, p. 342), "Homographic terms cannot be allowed in a retrieval system if selection of non-relevant information is to be avoided."

The next section discusses some methods for distinguishing between different types of homographs, and investigates whether any of these distinctions provide a classification useful for the identification of terms likely to cause problems in indexing and retrieval.

Classification of Ambiguous Terms

Ambiguous terms and hazardous terms

An ambiguous term is any word or phrase having two or more possible meanings. Ambiguity is a relative characteristic, in that the same orthographic term may be ambiguous or not, depending upon how it is viewed. Most terms as they exist in the language, that is, in isolation from any meaningful context, have more than one possible interpretation. However, in the context of a structured arrangement of words such as a document text or a thesaurus, the possible interpretations may be reduced to only one. The set of all possible meanings associated with a term in the language, independent of whether the term is ambiguous in the context of any given usage, is referred to here as the "inherent ambiguity" of that term.

Problems arise in automatic indexing systems because an inherently ambiguous term can have well-defined but different meanings in the thesaurus and in document usages. When this is the case, the

meaning of the textual passage and the meaning of the descriptor posted to it do not match, and the kind of error categorized here as Type-A occurs. Inherently ambiguous terms with the capability of causing Type-A errors are referred to in this paper as "hazardous" terms. Not all inherently ambiguous terms are hazardous in any given indexing environment. When the contextual definition of a thesaurus term indicates a range of acceptable meanings, and all of the possible meanings of the term in the language fall within this range, no errors of this sort can occur.

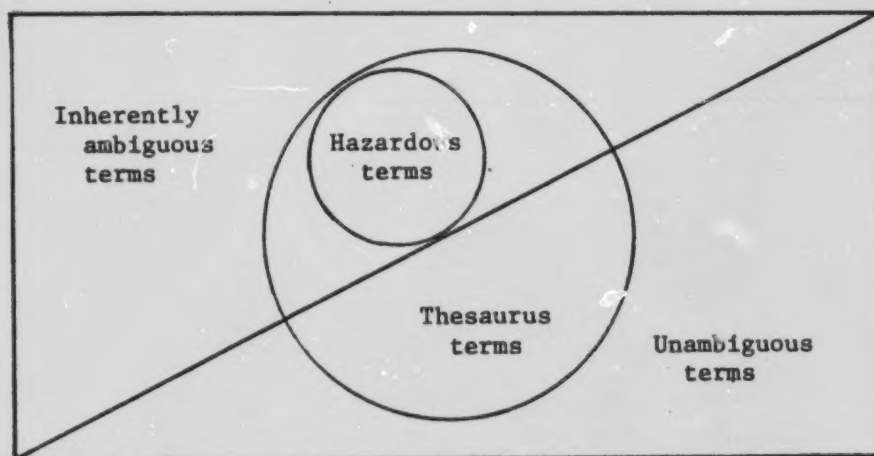


Fig. 2. Hazardous and inherently ambiguous terms

To indicate whether hazardous terms invariably caused indexing errors, a list of the hazardous terms in the "Master's Papers Thesaurus" was compiled from all of the index terms which actually occasioned the posting of one or more descriptors. Terms were selected if a usage could be imagined which was neither obsolete nor technical, yet was clearly not congruent with the thesaurus-defined meaning of the term. Development, for example, although entered under the hierarchical

division "Terms of general application," would not be a desired posting for the phrase 'housing development', and so was considered a hazardous term.

Of the 66 such terms identified, only 30 were found to have occasioned Type-A errors. These terms are listed in appendix D, in the immediate context of each of their occurrences in the document collection. The remaining 36 terms (see appendix E) were involved in a total of 252 term matches, yet were responsible for no errors related to term ambiguity.

It would be helpful in many phrases of information system design if there were some way to identify which hazardous terms were likely to cause problems in indexing and retrieval. This information could be used to guide thesaurus construction teams in the selection and hierarchical placement of index terms, and in deciding whether to tag descriptors with warning scope notes. Disambiguation efforts could be concentrated on the type of terms most likely to cause errors even before indexing failure occurs.

Classification according to the nature of ambiguity

Homographs and polysemes

One possible way of categorizing ambiguous terms is according to the nature of the ambiguity displayed. Linguists often make a formal distinction between homographs and polysemes. An orthographic word having two or more meanings with different derivations and no semantic components in common--e.g. 'pitcher' (the container) and 'pitcher' (the ballplayer)--is considered a homograph. In contrast, a polyseme is a "wordform that has two or more sememes [sets of semantic

components] having a number of common semantic components and differing in only a few . . ." (Hutchins 1975, p. 50). Following this distinction it is evident that words of common origin are more likely to be polysemic than formally homographic.

The hazardous terms in the "Master's Papers Thesaurus" were tentatively typed as either formally homographic or polysemic in accordance with the etymology given in the dictionary and were then checked to assure that all of the sememes involved were actually discrete. Any term with two or more unique sets of semantic components was categorized as a homograph, whether or not it also had meanings with sememes in common. Here, as in all other classifications discussed in this section, index terms not appearing as dictionary entries, and Economics (a different orthographic word than 'Economic') were excluded, leaving a list of 55 terms (appendix E, table 14).

Only four terms--Fine*, Grant*, Novel*, and Press*--proved formally homographic, and the latter two are questionable. Grant*, a homograph only on account of its use as a proper name, and Fine* were responsible for three errors out of seven total postings, while the other terms occasioned none.

The number of formal homographs discovered is too small to allow any conclusions regarding their likelihood of occasioning errors. However, concentrating on formal homographs alone would certainly ignore a large proportion of error-causing index terms.

True and contextual ambiguity

A different categorization, also based upon the inherent ambiguity of a term, is offered by Stephen Weiss (1973). Calling the

"semantic function" of a word the "influence that it exerts on its context," he defines "true ambiguity" as the case when two or more of the possible meanings of a term exert different semantic functions. "Contextual ambiguity," on the other hand, is the case when a word has a single semantic function, but different contextual interpretations. Weiss asserts that disambiguation efforts should concentrate on truly ambiguous terms, as these alone will have significant adverse effects on retrieval.

To some extent the difference between true and contextual ambiguity parallels the formal distinction between homographs and polysemes, for the contextually ambiguous term will generally be polysemic, having a wide range of meanings each sharing several sets of semantic components with the others. However, as Weiss defines it, true ambiguity does not require that the derivation of the orthographic words be distinct, or that the meanings have no common sememes. In the examples he uses, 'degree' (as in 'degrees Fahrenheit' and 'academic degree') is truly ambiguous, while 'base' (as in 'first base', 'military base', 'lamp base', and 'base register') is not. Both of the meanings of 'degree' stem from the Latin degradare and all share the sememe signifying step, rank or grade, while 'base' is derived from the Latin basis in the first three usages and from the Greek basson in the last.

Weiss' distinction leaves more room for interpretation than that between formal homographs and polysemes. For the latter, one need only determine whether the meanings associated with a word are entirely discrete. For the distinction between true and contextual ambiguity, one must decide at what point alternate meanings of a term

can be said to exert different semantic functions. Alternate interpretations of a term are rarely unrelated, and some point along a continuum of semantic variations must be chosen.

Two criteria for making this distinction were tried. For the first, most numbered definitions of a term in the Random House Dictionary of the English Language were considered indicative of different semantic functions. Exceptions involved definitions which read, "to _____," "the act of _____ing," "something _____ed," etc. These were considered to exert a single semantic function.

This method offered the advantages of ease and consistency, but it actually typed more variations in meaning as separate semantic functions than if Weiss's examples had been strictly followed, and the resulting division of terms was unsatisfactory. Only eight index terms of the 55 classified were contextually ambiguous by this standard, five of which occasioned errors (62.5%). The remaining terms divided about equally, with 42.5% occasioning errors.

A less objective but more realistic categorization of terms was therefore attempted. Terms for which all definitions could be reduced to one common factor were considered to exercise a single semantic function. 'Circulation', for example, means a moving around, whether of library books, blood, newspapers, etc., so the term was typed as contextually ambiguous despite the variety of definitions.

This classification showed the majority of terms to be contextually ambiguous, but still failed to differentiate between terms involved in errors and those which were not. The categorization of terms produced by both of these methods is presented in appendix E, table 15.

Whatever method is employed, the attempt to determine whether a greater percentage of truly ambiguous than contextually ambiguous terms cause indexing errors suffers from two fundamental problems. First, the categorization must consider all of the possible meanings associated with a term, while not all of these meanings are equally likely to occur in any given indexing environment. Second, the inherent ambiguity of a term does not change, while errors of commission as defined here are dependent upon the construction of the thesaurus, and will vary as the thesaurus is changed to impart different contextual definitions to the index terms involved.

A more realistic use of this distinction, then, would determine the type of ambiguity displayed based only upon a comparison between the document-contextual and thesaurus-defined meanings of a term for any given usage. This approach was taken in an attempt to find a correlation between the nature of a term's ambiguity and the frequency of errors committed by the term.

Differences in meaning between words in the text and thesaurus descriptors categorized as Type-A errors were examined to determine if true or contextual ambiguity was involved. The error was typed as related to true ambiguity if the different usages of the term were found to exercise different semantic functions, by the second standard employed above. Fifteen terms and 92 errors involved true ambiguity, while 12 terms and 28 errors involved contextual ambiguity. Table 3 summarizes the differences between these classes. Two index terms, Reader* and Reading*, were responsible for errors involving both types of ambiguity, so complementary percentages will not necessarily add up to 100%.

These figures indicate that Weiss might be justified in

TABLE 3.--Errors related to true and to contextual ambiguity

	True Ambiguity	Contextual Ambiguity
1) Number of index terms involved	15	20
2) Percentage of all 25 index terms	60%	48%
3) Number of errors	92	28
4) Average number of errors per term	6.13	2.33
5) Total number of postings	193	173
6) Ratio of errors to total postings	47.67	16.18
7) Ratio of errors to total errors	76.67	23.33

focusing on true ambiguities, for these seem to be involved in a disproportionate number of errors. However, a single index term, Literature*, was responsible for 57 errors, almost 62% of all errors involving true ambiguity. Literature* is atypical both in the frequency of its occurrence and in its high ratio of errors to total postings, and is clearly distorting the results in table 3 in favor of the contextually ambiguous terms. The results of the same calculations omitting Literature* are given in table 4.

While the terms involved in true ambiguity still appear to be responsible for more indexing errors, a chi-square test performed on the distribution of errors in (3) in table 4, using a distribution proportionate to the distribution of postings as the expected value, arrived at a chi-square of 3.404, which is not significant at the .05 confidence level for one degree of freedom. The difference between these figures is therefore not an indication that the distinction

between true and contextual ambiguity is meaningful in this respect, for it could be due entirely to chance.

TABLE 4.—Errors related to true and to contextual ambiguity--
Literature* omitted

	True Ambiguity	Contextual Ambiguity
1) Number of index terms involved	14	12
2) Percentage of all 25 index terms	58.33	50.00
3) Number of errors	35	28
4) Average number of errors per term	2.50	2.33
5) Total number of postings	136	173
6) Ratio of errors to total postings	25.74	16.18
7) Ratio of errors to total errors	55.56	44.44

The distinction between true and contextual ambiguity, then, does not serve to identify a subset of error-causing terms involved in a disproportionate number of indexing errors. It also fails to identify, from analysis of the inherent ambiguity of terms, a subset of hazardous index terms particularly likely to occasion errors. Attempts to objectify the distinction itself were only partially successful, and the categorization is tedious and difficult to perform consistently. Analysis according to the nature of ambiguity, then, does not seem to be a promising approach to the identification of problem terms.

Technical and non-technical terms

An alternate way of categorizing hazardous terms is according to whether they form part of the technical vocabulary of a discipline.

It is generally agreed that term ambiguity constitutes a greater problem in subject areas with a "soft," or relatively imprecise, vocabulary, than in those fields with a "hard," or technical terminology (Bhattacharyya 1974, Hines 1968, Walls 1972). This implies that technical terms are less ambiguous, or less prone to errors related to term ambiguity, than common terms used within a discipline.

To investigate whether technical and non-technical terms show different patterns of indexing failure, the hazardous thesaurus index terms were classed according to this distinction. The criterion for identifying technical terms was inclusion in either the ALA Glossary of Library Terms or the British Library Association's Librarian's Glossary. This provides only a rough standard, for the latest revised edition of the ALA work was published in 1943, while the Librarian's Glossary naturally favors British usage. However, these glossaries do provide a corpus of terms which can be considered "technical" in that they pertain exclusively to librarianship or have special meanings in library science or related fields.

Of all terms which committed errors relating to term ambiguity, 11 were categorized as technical according to this standard. These were responsible for 21 of the 133 Type-A errors committed. Table 5 indicates the average number of errors per term and the ratio of total Type-A errors to total postings for the technical and non-technical terms and for all error-causing terms as a whole.

These figures indicate that the technical terms are used in a more consistent manner than the other homographs. The difference in the ratio of total errors to total postings for the technical and non-technical terms was found to be statistically significant. A chi square

TABLE 5.--Commission of Type-A errors--Technical and non-technical terms

	Type-A errors per term	Type-A errors to total postings
Technical terms	1.91	.18
Non-technical terms	5.89	.48
All terms	4.43	.38

test performed using an expected distribution of errors proportionate to the distribution of terms shows $\chi^2 = 24.96$, with 3.81 significant at the .05 level. A second chi square using an expected value proportionate to the distribution of total postings arrived at $\chi^2 = 18.76$. Either these terms are less ambiguous, i.e., possess fewer alternative meanings, then the other error-causing terms, or the primary (thesaurus-defined) meaning is simply more prevalent in normal usage.

However, this does not necessarily imply that proportionately fewer technical terms will be involved in indexing errors than non-technical terms. If this were so, the distinction would provide a simple criterion for the identification of relatively "safe" hazardous terms.

To investigate this possibility, it was necessary to determine whether the technical terms were represented among the error-causing terms in proportion to their frequency in the thesaurus as a whole. To indicate what proportion of the thesaurus vocabulary could be considered technical, the 100 index terms in the random sample taken of all thesaurus index terms as described in part one, were checked for inclusion in either glossary. Results are noted in appendix C.

Exactly one-fourth matched entries verbatim. Several multi-word index terms, however, differed only slightly from phrases included in the glossaries, or were variant forms of descriptors which did appear as entries, e.g. Biographical and I&R, non-print synonyms for Biography and Information and referral services, respectively. Including these near-matches in the count brings the percentage of technical terms in the sample up to 31%, or nearly a third.

Fifteen of these terms, or 48%, actually occasioned indexing, as compared with 41% of the non-technical terms. Assuming that the average number of postings per term is not significantly different for technical and non-technical terms, the number of technical terms expected to cause Type-A errors if there is no difference in behavior between the two groups can be estimated, as shown below:

	Number of terms in thesaurus	Percentage of terms indexing	Estimated opportunities for error
Technical	512	.48	246
Non-technical	1138	.41	467

Estimated number of technical terms occasioning errors
if no difference in behavior between the two types:

$$\frac{246}{713} \times 30 = 10.35$$

This figure is almost identical to the number of technical terms actually found to have occasioned errors relating to term ambiguity (11). It indicates that the distinction between technical and non-technical terms does not identify a class of hazardous terms disproportionately unlikely to occasion one or more indexing errors.

In conclusion, attempts to find a meaningful correlation between some quality inherent in the term in the language and a suggestive pattern of indexing failure were generally unsuccessful. An alternate approach to the errors committed takes the point of view of what can be done to prevent them. The following section discusses various methods of disambiguation, and investigates the application of some of these techniques to the indexing performed on the master's papers collection.

Disambiguation of Homographs

Most of the literature on the disambiguation of homographs pertains to thesauri designed for manual indexing. Here the solutions are relatively easy to implement, and authorities differ mainly on details of format.

The COSATI Guidelines for the Development of Information Retrieval Thesauri (1967, p. 2) recommends three methods for avoiding ambiguities: (1) using compound (precoordinated) descriptors instead of single word terms for greater specificity, (2) providing short explanations or definitions in scope notes, or (3) adding parenthetical qualifiers to the descriptor.

Aitchison and Gilchrist's Thesaurus Construction suggests that parenthetical qualifiers be added only to the less frequently used senses of a term (1972, p. 18). Dagobert Soergel recommends not only the addition of qualifying expressions, but also the numbering of each homographic descriptor (1974, p. 301).

Following each recommended format in turn, then, we might have:

Pitchers (baseball)
Pitchers (pottery)

Pitchers
Pitchers (pottery)

Pitchers-1 (baseball)
Pitchers-2 (pottery)

The basic idea, however, is the same.

Obviously, when indexing is to be performed automatically, the nature of the problem is quite different. The aim is not to clarify the meaning of the term within the thesaurus, but to prevent homographic terms from occasioning the postings of inappropriate descriptors, or to ameliorate the detrimental effects of these indexing errors on retrieval performance.

If the indexing is to be used in a post-coordinate retrieval system, the combination of a homograph with other subject terms can serve to screen out documents with inappropriate term usages. In a Boolean environment, for example, a user interested in baseball pitchers might submit a request for documents containing both 'sports' and 'pitchers', as well as any number of other terms. Only documents concerning the desired variety of pitcher are likely to meet these requirements. Term ambiguity will be a problem when usages co-occur in irrelevant or unanticipated relations ('sports' and 'pitchers' might, for example, retrieve a document on golfing irons), or when the user, opting for recall, fails to specify co-occurrence requirements.

In an experimental non-Boolean retrieval system, called SMART, these problems are ameliorated. A document (text or query) in this system is described by a single vector in document space, which is determined from calculations based on all of the indexed terms in the document. The more similar the bulk of the vocabulary, the closer the documents will be spaced. Documents concerning baseball pitchers, then,

are likely to show high correlations with each other whether or not they contain the particular word 'sports', or whatever term the user might think of specifying in a Boolean query.

This natural tendency can be improved upon with the use of appropriate thesaurus-mapping algorithms. According to Gerard Salton (1968, p. 24), "Even if almost all terms used in a given context are inherently ambiguous, the juxtaposition of many multiple mappings can often identify the appropriate concept classes with reasonable accuracy. The relevant categories will normally be reinforced, since they apply to many terms, whereas the extraneous categories will be randomly distributed."

A semi-automatic method for disambiguation utilizes multiple postings in conjunction with a rather broad subject classification. Human catalogers assign class symbols to each document before automatic indexing is performed, and the class number of the document is automatically added to each index term posted to it. 'Pitcher', then, might post to one document in combination with the class symbol for sports, and to another with the symbol for crafts. This method could be helpful in multi-disciplinary information systems, as many words which are relatively well-defined within a particular subject area become ambiguous in a more heterogeneous environment. One information scientist, for example, reported that she had actually compiled a thesaurus of library science terminology from words included in the Engineers Joint Council's Thesaurus of Engineering Terms (Richmond 1973). These words would complicate any attempt to make thesauri for the two fields compatible with each other.

Precoordination is another approach to reducing the ambiguity

of index terms. The descriptor Computer programs, for example, will not match the homograph 'programs' used in reference to television shows or syllabi. Post-coordinating 'computer' and 'program' in retrieval will achieve much the same effect but with a greater likelihood of false drops. When indexing is performed automatically, however, the gain in precision achieved through precoordination will be accompanied by a loss to recall; unmodified occurrences of 'program' will not index even when computer programs are clearly implied.

Other techniques which have been tested involve the indexing process rather than the thesaurus index terms. One method employs algorithms for searching the text contiguous with the homograph for key words or "clues" that indicate the usage of the term. Rules for identifying each clue and the concept class to assign if it is or is not found are stored for every homograph for which disambiguation will be attempted (Stone 1969; Weiss 1973). A similar procedure, devised for application in machine translation, looks for the word or phrase determiner in the sentence based upon "automatically recognizable statistical features of context" (Swanson 1961, p. 391).

Context may also be utilized in automatic syntactic analysis. In its most basic application, words which can function only as one part of speech are identified (usually by means of a dictionary lookup, although other techniques exist) and tagged as such. Grammatically ambiguous terms can then be disambiguated by checking the patterns formed by each of the possible grammatical functions of the term in context against a dictionary of allowable phrase structures. A noun-verb ambiguity, for example, is resolved if the possible structures are found to be ARTICLE-NOUN or ARTICLE-VERB. ARTICLE-NOUN is

the only acceptable combination, and the homograph is tagged as a noun (Bobrow 1967). Where the meaning of a homograph is entirely dependent upon its grammatical function, as is the case, for example, with the term 'objective', syntactic processing could be applied to the disambiguation of index terms.

The remainder of this section takes a closer look at some of these techniques--syntactic tagging, precoordination, and multiple-posting in conjunction with post-coordination--and investigates their application to the ambiguous index terms in the "Master's Papers Thesaurus." First, however, we will examine the effects of a relatively simple procedure, that of requiring a truly exact match between index terms and text before indexing can occur.

Exact match

In the Automatic Indexing system used in this project, the thesaurus file and document file are each submitted to normalizing programs before any matching is actually performed. In these preliminary programs the text is capitalized, punctuation marks and stop words are deleted, and plural words are converted to singular form. Therefore, a truly identical match between the thesaurus index term and a portion of the text is never required for a match to be found, even under the strictest of the phrase match criteria. For example, 'Fine' and 'fines'; 'ILL', 'Ill.', and 'ill', are all found to match exactly with each other. Were it possible to allow indexing to occur only when a truly identical match were found between text and index term, a number of the errors which occurred in the indexing of the 149 master's papers could be prevented.

Distinction between singular
and plural word forms

Of the 60 hazardous index terms identified in the "Master's Papers Thesaurus," eight terms would be affected by changes in the suffixing algorithm. These terms, and their actual and potential usages, are listed in table 6.

Four index terms--Economics, Fines, Objectives, and Westerns--were responsible for errors because the adjectival forms of the words are indistinguishable from the singular nouns. (Economics constitutes a special case, being a plural noun construed as singular.) The meanings of these terms as adjectives are so different from the nounal thesaurus-defined meanings that adjectival usages in the text will invariably occasion errors unless indexing can be prevented. As the adjectival forms will never have the -s suffix, this could be accomplished by allowing the term in the text to occasion a posting only if it matched exactly the plural thesaurus term.

However, while this would effectively exclude the indexing of all adjectival usages, it would also eliminate the desired indexing of the singular noun form for all of these terms except Economics. All of the noun usages of these terms in the document collection were plural, indicating that the singular word-forms may be less common, but there is no reason why they might not be used in future additions to the collection. Of the hazardous terms which did not cause errors only one fits into this category: Novels, a lead-in term for the descriptor Fiction. This term was used only as a plural noun, but could conceivably occur as a singular noun or adjective.

For four other index terms, the presence or absence of the plural suffix could help distinguish between acceptable and

TABLE 6.--Singular and plural word forms--actual and possible use

Term stem	Possible Use		Actual Use				
	Singular ending (-, -e, -y)	Plural suffix (-s, -es, -ies)	Singular ending (-, -e, -y)	Plural suffix (-s, -es, -ies)	Precision	Matches: no. correct/ no. errors	Matches: exact match only
Economic	E	S	E(4)	-	0	0-4	0
Facilit	S	C	-	C(5)	1.0	5-0	5
Fine	S	C	E(1)	C(3)	.75	3-1	3
Humanit	E	C	-	C(2)	1.0	2-0	2
Novel	S	C	-	C(5)	1.0	5-0	5
Objective	S	C	E(1)	C(11)	.92	11-1	11
Publication	S	C	E(3) C(5)	C(10)	.83	15-3	10
Reading	S	E	E(4) C(8)	E(2)	.57	8-6	12
Western	S	C	E(1)	-	0	0-1	0

KEY TO SYMBOLS: Possible use: C = Always will be correct
S = Sometimes will be correct
E = Always will be error

Actual use: C = Correct match
E = Error match
- = No matches

TABLE 6.--Continued

NOTE: Precision = total number of matches/total correct matches. With exact match only, precision for terms except for Reading is 1.0. Precision for Reading is .66.

All Type-B and Type-C errors excluded from count.

unacceptable noun meanings. Humanities and Publications will always be acceptable in the plural, although Publications may also occur in the thesaurus-defined sense in the singular. 'Facilities', which did not occasion any errors, parallels Publications in this respect. Reading will always occasion errors when used with the -s ending.

If exact match were required for only the "safe" terms Economics, Humanities, and Reading, six errors would be eliminated with no possible loss to recall. If it were to be required for all of these terms, 12 errors would be prevented at the expense of the omission of five desired postings, with the possibility of even greater loss to recall as new documents are added to the data base.

The distinction between singular and plural word forms could be achieved with the Automatic Indexing system by simply omitting the preliminary suffixing procedure. To do so, however, would require the addition of a non-print synonym to every nounal index term in the thesaurus for which matches to both singular and plural forms were desired, a time-consuming and expensive endeavor. An alternative would be to exempt some terms from the conversion and not others. Conversion in the system used here is performed by means of an algorithm which determines whether words of length 4 or greater should be put into their singular according to rules of the nature "ies→y," "es→e," etc. It would be necessary to store a dictionary of terms to be exempt from conversion and precede suffixing by a dictionary look-up. This, of course, would entail the expense of additional processing time and storage space.

In sum, it would be possible to recognize the difference between singular and plural forms for some terms, where this

distinction is enough to discriminate between acceptable and unacceptable term usages. If it appeared from this analysis that a large percentage of errors could be prevented in this manner, the trade-offs between costs and performance could be investigated for these and other alternatives. However, as only between three and nine terms and about 7% of errors are involved in this discussion, it seems as though this solution is not worth pursuing.

Distinction between upper
and lower case letters

Three descriptors posted in error a total of seven times due to the incapacity of the matching program to distinguish between capital and lower case letters. AMERICAN LIBRARIES and LIBRARY LITERATURE appear in the thesaurus as proper names of professional publications, but the phrases are used in the documents to refer to libraries in America and library science literature as well as to the periodicals. Instructions for inputting data specify that titles which would otherwise be underlined must appear in all upper case letters. Acronyms can also be identified through capitalization. If this distinction could be recognized, this source of ambiguity would be eliminated with no possible loss to recall.

At least two other descriptors, DIALOG and LIBRARY JOURNAL, are potentially ambiguous in this way. Moreover, of the 221 periodicals indexed in Library Literature, at least 46 titles not presently included in the thesaurus are subject to the same sort of ambiguity. Some facility for the recognition of titles and acronyms would therefore appear to be desirable. An alternative to extensive program modification could be a simple change in the rules for inputting

thesaurus terms and master's papers data. All titles and acronyms could be preceded, or preferably, followed by an unprintable character which would not be stripped before matching.

The third term in this category, Grants, was not used in the thesaurus as a proper noun but appears twice in the text as part of a name. As 'Grant' is a common surname, this sort of error can be expected to recur. However, as the initial letter of the word is the only clue to its usage, and as this letter will always be capitalized in sentence first position, even the recognition of upper case letters would not consistently disambiguate this term. There are no other index terms demonstrating a problem of this nature.

Syntactic tagging

Procedures for identifying grammatical parts of speech are utilized in many information storage and retrieval systems (Climenson, Hardwick, and Jacobson 1961; Fishman 1972; Klingbiel 1973). Codes representing grammatical categories may be stored along with the other information pertaining to an index term, and can be used for a variety of purposes. Some stemming techniques employ dictionaries of syntactically coded word stems and affixes, and must not only find matching roots and affixes for each term in the text, but also find an allowable match between the syntactic codes before "stemming" will take place. Grammatical relationships between words may be used in phrase recognition procedures for the automatic construction of phrase dictionaries. Simple dependency relationships or more precise syntactical relationships may be specified in retrieval, to increase precision.

In operational systems, syntactic tags are generally assigned by human indexers--the machine does not determine part of speech, but

rather utilizes this information to some end. Grammatically ambiguous terms may receive two or more syntactic codes and they often reduce the effectiveness of the system. Some work has been done in constructing algorithms for machine disambiguation of these terms, but grammatical ambiguity remains troublesome.

In this discussion, "part of speech" will generally refer simply to noun, verb, and adjective classes, leaving aside finer distinctions between verb tenses, gerund and participle functions, etc. Of the 30 index terms which committed errors relating to term ambiguity, 11 may be used as more than one part of speech, leaving aside for the moment the capacity of most nouns to function as adherent nouns (adjectivally). For three of these terms--Contemporary, Records, Grants--the semantic ambiguity is not tied to the grammatical ambiguity. For the remaining eight, the grammatical category alone is enough to distinguish between acceptable and unacceptable meanings of the term. Seven of these index terms were responsible for a total of 11 actual indexing errors.

Noun-adjective ambiguity

Four of the index terms responsible for Type-A errors--Westerns, Objectives, Fines, and Economics--have semantic ambiguity tied to the noun-adjective distinction. All are used as nouns in the thesaurus. Were it possible to discriminate between the occurrences of these word stems according to grammatical category, a total of seven errors, or about 3% of total errors of commission, could have been prevented. However, potential usages of the root terms must also be considered. Western*, Fine* and Economic* could conceivably be used adjectivally in senses appropriate to their respective concept classes (e.g.

'western novels', 'fine policies', 'economic theory', etc.) and these usages would have to be anticipated and added to the concept class as synonyms for indexing to be effected.

Of the hazardous terms which did not cause errors, only two root terms show true semantic ambiguity tied to the noun-adjective distinction--Novel* and Abstract*. Both were used in the documents only as plural nouns, and so occasioned no errors. (A third term, Colonial, like Contemporary, has semantic ambiguity which transverses the noun-adjective distinction.)

Adherent nouns represent a special problem. The majority of nouns in the English language may be used to modify other nouns. In general, the referent of the resulting phrase does concern the nounal thesaurus descriptor to such an extent that their indexing does not occasion errors. In other words, 'student profiles' do concern students, 'network facilities' concern networks, etc. Only two indexing errors were declared on account of adherent nouns, both involving the descriptor Student*. 'Student mail' and 'student workbook' were two textual usages considered, in context, inappropriate to the thesaurus-defined meaning of the terms.

Noun-verb ambiguity

A term which can function as both a noun and a verb is not generally a hazardous term. Simple stem dictionaries improve performance in part by grouping noun and verb forms into single concept classes. The thesaurus achieves the same effect for some descriptors by including verb or participle forms as synonyms to the main term. Grammatical ambiguity, then, is only a potential source of error in two cases: (1) when there is a substantial difference of meaning

between the noun and verb usages, and (2) when the thesaurus itself makes a distinction between them by assigning the different grammatical forms to separate concept classes.

Only one of the index terms involved in Type-A errors fits into our first category above: Aims, which appears in the thesaurus as a synonym for Objectives in the "Organization and administration" division of the hierarchy. 'Aim' as a noun has several dictionary definitions, only some of which are acceptable, but the meaning of the verb, "to strive, try," is considerably broader than the restricted management sense implied by the thesaurus. Both times 'aim' was used as a verb in the document collection it occasioned indexing errors.

The only other hazardous term in this category is Slides, for which the only definition as a verb is substantially different from the nounal thesaurus meaning as a non-book library material.

Thesaurus-defined distinctions involve the hazardous terms Catalogs and Indexes, which occasioned indexing errors, and Abstracts, which did not. While the user seeking documents pertaining to the object will not necessarily be interested in the process, the semantic overlap is sufficient that a false drop obtained by the matching of noun and verb usages has a good possibility of being relevant. A test collection of 30 queries representing users' information needs was indexed using the "Mater's Papers Thesaurus," and the only query containing either term requested information of the process of cataloging as well as on library catalogs. On the other hand, co-occurrence of these terms in the documents is rather low. 'Indexes' and 'indexing' co-occurred in only one document out of the nine which contained either one term or the other. The 15 documents indexed with either Catalogs

or Cataloging showed no co-occurrence. 'Abstracts' occurred in four documents, and 'abstracting' in none, except as part of a phrase indexed by a precoordinated descriptor.

In any case, differential treatment of the noun and verb usages would eliminate only two errors out of a total of 36 postings, with dubious benefits to real precision. An alternative approach might consider whether the distinctions between process and product made by the thesaurus are actually relevant to user needs.

Summary--syntactic tagging

Table 7 lists all of the index terms for which grammatical ambiguity is tied to semantic ambiguity, and compares actual indexing performance with the possible performance if these terms were tagged for part of speech and allowed to match only if thesaurus and document usage coincided. While the 11 errors which could be prevented represent a fairly large proportion of the 68 total occurrences of these terms in the document collection, they comprise only about 9% of the total errors due to term ambiguity, and only 4.8% of all errors identified. The nine terms involved represent only .55% of all thesaurus index terms. Grammatical ambiguity, then, does not seem to be a major cause of indexing errors. Considering that procedures for machine resolution of syntactical ambiguity are quite complex and not entirely successful, the conclusions from this thesaurus-based automatic indexing experiment indicate that the resultant improvement in performance would not be worth the effort involved.

TABLE 7.--Syntactic tagging--Error-causing terms

Index term (stem)	Thesaurus usage	Document usage (frequency)	Matches: correct- errors	Matches: same grammatical form	Actual pre- cision
Abstract	N	N(6)	6-0	6-0	1.00
Aim	N	N(1) V(2)	1-2	1-0	.33
Catalog ^a	N	N(19)	19-1	19-0	.95
Economic	N	A(4)	0-4	0-0	0.00
Fine	N	A(1)	0-1	0-0	0.00
Index	N	N(15) V(1)	15-1	15-0	.94
Novel	N	N(5)	5-0	5-0	1.00
Objective	N	N(11) A(1)	11-1	11-0	.92
Western	N	A(1)	0-1	0-0	0.00

Key to abbreviations:

N = Noun

V = Verb

A = Adjective

NOTE: Actual precision = total correct matches/total matches.
Precision for matches of same grammatical form only is 1.0 for all terms.

All Type-B and Type-C errors excluded from count.

^aIncludes 'catalogue'.

Precoordination

Precoordination, precision and recall

Of the 100 index terms in the random sample, 65 were pre-coordinated, if a precoordinated index term is defined as any index term consisting of two or more discrete words. The majority of these index terms were short phrases of two to four words which could reasonably be expected to appear together, such as 'reluctant reader' or 'information storage and retrieval'. Six were proper names of people or places.

If precoordination generally reduces recall, one would expect to find fewer precoordinated index terms actually scoring matches, in proportion to their numbers, than single-word index terms, or "uniterms." One would also expect that the precoordinated index terms which did post would score proportionately fewer hits per term than the uniterms. These expectations are borne out by the results of the sample.

Forty-two index terms in the random sample (42%) actually "hit," or matched a word or phrase in a document, at least once. Of the 35 uniterms, 17 actually hit, or 48.57%, while only 38.46% of the precoordinated index terms hit. Furthermore, average hits/term numbered 3.88 for the uniterms as opposed to 2.40 for the precoordinated index terms. Results are summarized in table 8.

The precoordinated terms in the sample, therefore, did match fewer terms in the document collection than did the single-word index terms. It would also be expected that the use of precoordinated index terms would lead to greater precision in retrieval than the post-coordination of uniterms. To verify this, all of the precoordinated

TABLE 8.--Precoordination and term matches--Random sample index terms

	Number terms in sample	Number terms that hit	Percentage of terms that hit
All index terms	100	42	42.00
Uniterms	35	17	48.57
Precoordinated index terms	65	25	38.46
	Number terms that hit	Total number of hits	Average number hits/term
All index terms	42	126	3.00
Uniterms	17	66	3.88
Precoordinated index terms	25	60	2.40

terms in the thesaurus that were involved in at least one match were noted, and the total number of hits and errors recorded for each. The precoordinated index terms were then broken down into their component uniterms, and retrieval combining the uniterms in each phrase with a Boolean AND was simulated. A document "retrieved" in this manner was considered "relevant" if the meaning of the precoordinated index term was appropriate to it. The details of this process are presented fully in appendix F, table 16.

The results of this investigation indicate that the use of post-coordinated uniterms would retrieve approximately twice as many relevant documents as the precoordinated index terms would, but that both micro- and macro-precision would be substantially reduced. The precoordinated terms do show extremely high precision. However, all

of the documents retrieved by the precoordinated terms would also have been retrieved by the uniterms, along with an additional seven relevant documents, an improvement in performance of about 32%.

While it is clearly impossible to compare precision measures for the index terms which did not index, it is worth considering whether the use of post-coordinated single-word terms would retrieve a slew of irrelevant documents, a substantial number of relevant ones, or nothing at all. Therefore, a similar experiment was conducted using the uniterm components of the precoordinated index terms which found no matches in the document collection (see appendix F, table 17). Results show that Boolean retrieval requiring the co-occurrence of the individual words would have produced five relevant documents and three irrelevant ones, yielding a respectable micro-precision measure (.625). The precoordinated index terms of course, never posted at all, and would have retrieved nothing.

Most of the uniterm components of precoordinated index terms were found to have occurred in the document collection with relatively high frequency, especially when compared with the precoordinated terms themselves. Matching, storing, and combining these terms would require additional effort, resulting in greater operating time or expense as the price for additional recall. Whether or not this expense is worthwhile depends upon the precision and recall requirements of the user group.

Precoordination to resolve ambiguity

An investigation into the possible precoordination of ambiguous index terms touches upon the manner of error classification employed in this study. A Type-B error, related to flaws in the thesaurus, was

designated whenever the omission of an index term from the appropriate concept class resulted in the posting of a less appropriate descriptor. An example of this involves the descriptor Pictures, used in the thesaurus to refer to picture reference collections. A second descriptor, Films, is obviously the appropriate index term for the phrase 'motion pictures'. Each time the occurrence of 'motion pictures' in the text occasioned the posting of Pictures, therefore, this was not categorized as a Type-A error due to the ambiguity of the word 'pictures', but rather as a Type-B error, or one due to the omission of a desirable synonym in the construction of the thesaurus.

Because of this method of error classification, and also because of a preference shown to precoordinated forms in the "Master's Papers Thesaurus," we would not expect to find many of the terms responsible for errors of ambiguity capable of disambiguation through precoordination. While one can nearly always think of some phrase in which to imbed an ambiguous term, the primary consideration is whether the proposed combination of words will occur in the language of the documents frequently enough to merit inclusion as a precoordinated index term. This cannot be predicted with perfect accuracy, but past term usage does tend to indicate probable future use.

The context of all term occurrences associated with Type-A errors was therefore examined. Appendix D shows these usages in context up to the first stop word or punctuation mark. This listing shows that for most of the terms, correctly indexed occurrences do not appear in any consistently repeated context suitable for inclusion as a pre-coordinated form. Record*, for example, appears four times out of eight with 'reviewing', but the meaning of 'record reviewing' is too

specific to be used for the generic concept "records." Objective* occurs with 'aims' and/or 'goals' four times of 11, but as both words belong to the same concept class as Objective*, simply eliminating Objective* from the class would produce the same effect more simply than attempting to precoordinate the term. Either alternative would lead to a substantial drop in recall. The only true candidate for precoordination is Fine*, which occurs in two correct usages out of three as 'library fines'. Precoordinating the descriptor here would reduce recall by a third but serve to screen out most adjectival usages of the word. If high precision were required this could be considered, although little can be concluded from such a small sample of textual usage.

The idea of employing precoordination to prevent undesired term matches appears to offer more promise. There is a facility in the indexing system which allows words or phrases to be excluded from the indexing process. When a match is found to a specially designated thesaurus term, the descriptor for the concept class of which the term is a member will not be posted. If a particular word combination were found to be consistently associated with desired term meanings, then, the phrase could be entered into the thesaurus as one of these posting-prohibiting terms.

'Aim', for example, is followed in both improperly indexed occurrences by 'to', and the dictionary confirms that the verb 'aim' is generally followed by the prepositions 'to' or 'at'. Including 'aim to' and 'aim at' as posting-prohibiting terms in the concept class for Aim* would improve the precision of the indexing from .33 to 1 at no expense to recall for the documents in this collection.

'Classification' is preceded by 'Factor Analysis' twice and 'personnel' once. 'Factor Analysis' will probably prove an uncommon modifier, but 'personnel classification' might occur often enough to be considered as a posting-prohibiting term.

'Climate' is followed twice by 'of', but prohibiting indexing when 'climate of' is matched would exclude such possible acceptable usages as in "the climate of equatorial Brazil." The same circumstance applies to 'publication of'. 'Series of', however, might make a good discriminator, for in library literature 'series books' is a far more common usage than 'series of books' or similar constructions. Given this document collection, excluding 'series of' from indexing would eliminate two errors with no loss to recall.

A final possibility is the addition of 'the literature' as a posting-prohibiting member of the concept class Literature. While the definite article could occasionally precede an acceptable use of the term, it is undoubtedly more common in reference to "the literature" of a discipline, an unacceptable usage. With this collection, prohibiting the posting of Literature when 'the literature' is matched would eliminate 26 errors, while allowing all acceptable usages to be indexed. It would also be possible to include 'X literature' as a non-print synonym for all Xs, where X is a discipline listed in the thesaurus division "Types of literature." Medicine, for example, could be given the non-print synonym 'medical literature'. Using both of these techniques would increase precision from .05 to .19 for Literature at no loss to recall for this document collection. Only 16 of the 57 errors committed by Literature appear unpreventable by any means.

In conclusion, only five index terms have precoordinated forms

suitable for inclusion, either as regular synonymous terms or as special posting-prohibiting terms. The latter application appears to be more promising. The addition of posting-prohibiting index terms to the concept classes for Literature, Aim*, Series and Classification would eliminate 31 errors of the 75 committed, while preventing none of the desired instances from indexing. However, the document collection indexed for this analysis is very small, and the possibility exists that new documents will show quite different patterns of word usage. A second difficulty is that the indexing system would have to be capable of a differential treatment of stop words for the posting-prohibiting facility to be employed for the terms 'aim to', 'series of' and 'the literature'. Even so, exploitation of this facility merits consideration as a relatively simple method for increasing the precision of indexing.

Post-coordination

Most operational automatic indexing and retrieval systems rely upon post-coordination for the resolution of ambiguities. With a hierarchical or faceted thesaurus this means that the semantically ambiguous term must appear in all appropriate areas of the thesaurus. An example of double-posting in the "Master's Papers Thesaurus" which was included accidentally and later corrected, involves the index term ILL. ILL was entered as a non-print synonym for Interlibrary loan, which appears in the "Public services" division of the hierarchy. 'Ill', which is treated in the matching process as if it were identical to ILL, appeared as a non-print synonym for Illinois in the "Geographical list." Each occurrence of these abbreviations in the text (and,

incidentally, also the word 'ill') caused the posting of two descriptors, Illinois and Interlibrary loan.

In many systems multiple-posting is intentional and systematic. Irrelevant usages of homographs are hopefully screened out in retrieval. If term weights are used, the weight of the homograph may be distributed between all concept classes in which the term is included. In our example, if the standard term weight were 12, Illinois and Interlibrary loan could both post with weight 6.

One cannot predict very accurately how effectively the undesired meanings of the potentially ambiguous terms would be screened out if multiple postings were allowed, for the efficacy of this strategy depends upon the retrieval, not the indexing, process. We can, however, investigate how well multiple postings would fit into the structure of the thesaurus, for if an ambiguous term is to occasion the posting of more than one concept class, these classes must be available in the thesaurus.

For any term with multiple meanings at least one of the following situations must apply: (1) the term could post appropriately to two or more concept classes already included in the thesaurus, (2) the term could have a meaning that is unprovided for in the thesaurus, but would be a logical and desirable addition, and (3) the term could have meanings that it would not be desirable to index. Obviously, only terms in the first two categories are candidates for multiple-posting.

When the terms occasioning Type-A errors were categorized according to this scheme, 14 were found to be unsuitable for multiple-posting. For example, the thesaurus generally eschews uniterm adjectives, so representation of the adjectival meanings of 'fine' and

'objective' would not be consistent. Other terms, such as 'acquisition(s)', 'series' and 'selection', do not appear to be very useful in any but their specific library interpretations.

Of the 16 candidate terms (shown in table 9) all but one had two or more meanings which could be accommodated by the thesaurus as it now stands. Western*, used in the thesaurus to designate a literary genre, would require the addition of a new concept class in the hierarchical division "Geographical list" to allow for its other sense to be properly indexed. (A third index term, Western, to designate the television programs, did not seem necessary.)

When a single term is allowed to cause the posting of two unrelated descriptors, of course, one of the descriptors will invariably occasion an error related to term ambiguity. If the index term was originally assigned to the thesaurus concept class representing its most frequent usage, multiple-posting will actually increase the number of errors committed. In this case, if the candidate terms were allowed to post to all concept classes appropriate to them as shown in table 9, the ratio of total correct postings to total Type-A errors would be .84. This compares with a ratio of .75 for the indexing actually performed. This does not, however, indicate that multiple-posting would not improve performance in retrieval.

The hierarchical arrangement of a thesaurus is based upon semantic relationships between terms, and in principle, the more closely two terms are related, the nearer they should appear to each other in the hierarchical structure. In addition, related terms have been shown to co-occur with relatively high frequency in documents (Salton 1968, p. 118). The employment of post-coordination to exclude

TABLE 9.--Multiple-posting of candidate terms

Candidate	Concept Classes			
	Descriptor	Eldest ancestor ^a	Division [†]	Correct matches-errors
Bibliography	*Bibliography	Books	C	19-2
	Bibliographers	Personnel	D	2-19
Catalogs	*Catalogs	Bibliographic control	F	6-4
	Cataloging	Bibliographic control	F	1-9
Classification	*Classification	Bibliographic control	F	2-4
	Job description	Staffing	D	2-2
Dictionary	*Dictionary	Books	C	0-1
	Authority file	Bibliographic control	F	1-0
Economics	*Economics	Types of literature	A	0-4
	Economic factors		K	4-0
Indexes	*Indexes	Books	C	15-1
	Indexing	Bibliographic control	F	1-15
Journals	*Periodicals	Serials	C	10-1
	Manuscripts	Special collections	C	1-10
Library building	*Library construction	Buildings and grounds	D	1-2
	Library architecture	Buildings and grounds	D	2-1
LIBRARY LITERATURE	*LIBRARY LITERATURE	Books	C	1-4
	^b Library education		A	4-1
	National organizations	Professional journals	A	4-1

TABLE 9.—Continued

Candidate	Concept Classes			
	Descriptor	Eldest ancestor	Division [†]	Correct matches-errors
Literature	*Literature	Types of literature	A	3-57
	^c Types of literature		A	19-41
	Publications	Communications	A	57-3
Philosophy	*Philosophy	Types of literature	A	0-1
	Attitudes		K	1-0
Publications	*Publications	Communications	A	15-3
	Publishing	Book trade	B	3-15
Reading	*Reading	Library and society	A	8-6
	^d Books	Library materials	C	4-10
Records	*Records	Non-book materials	C	8-13
	Government publications	Special collections	C	3-18
	Archival materials	Special collections	C	5-16
	Statistics		K	6-15
Subject analysis	*Subject analysis	Bibliographic control	F	0-1
	Collection evaluation	Collection management	F	1-0
Westerns	*Westerns	Types of literature	A	0-1
	^e West		G	1-0

NOTE: Terms marked with an asterisk (*) represent the class to which the candidate is actually assigned in the thesaurus.

^aIf there are no broader terms between the descriptor and its division label, no "eldest ancestor" is listed.

TABLE 9.--Continued

^bFollowing a proposal offered in Section 3-2 of this paper, 'library literature' should be an acceptable synonym for 'library science'. Unfortunately, 'library science' appears in the thesaurus only as a synonym for 'library education'. As 'professional journals' is more specific than 'library literature', posting to this concept class would decrease precision.

^cPosting both father and son concept classes here effectively renders the son ('literature') superfluous.

^dThe concept class designated by the description 'books' is actually broader than the term 'books' itself. One non-print synonym in this class is 'printed materials'.

^eA proposed descriptor not actually in the thesaurus.

[†]Division codes are as follows:

- A = General
- B = Bookmaking, publishing and the book trade
- C = Library materials
- F = Technical services
- G = Geographical list
- K = Terms of general application

undesired usages of homographs, on the other hand, presupposes that the various meanings, and hence the hierarchical locations, of the homographs will be far enough apart that each meaning has high correlation coefficients with different sets of terms. One way of checking on the effectiveness of this technique, therefore, is to see how closely the different concept classes posting from a single term occurrence are related to each other in the thesaurus.

Concept classes in the same major division of the hierarchy ought to be more closely related than classes in different branches, and concept classes with the same parent ought to be more closely related than those in the same branch but having different parent nodes. Assuming that the candidate terms are allowed to post to each of the concept classes indicated in table 9, we can divide the terms into three classes as shown below:

1. Terms belonging to two or more concept classes appearing in the same hierarchical division, with the same highest parental node--Catalog*, Library building, Literature, Record*

2. Terms belonging to two or more concept classes in the same hierarchical division, no common parent node--Journals, LIBRARY LITERATURE, Subject analysis

3. Terms belonging to hierarchically unrelated concept classes--Bibliography, Classification, Dictionary, Economic*, Index*, Philosophy, Publication*, Reading, Western*

Retrieval should in general be most effective for index terms in the final category, and least effective for those terms in the first.

Judging from this rough indicator, then, we could expect that

if these index terms were allowed multiple postings, semantic discrimination by co-occurrence requirements in retrieval would be poor for about 25% of them, medium for about 19%, and fairly good for about 56%. If we add the poor discriminators to the set of terms for which multiple postings would be impractical or undesirable, we can arrive at a final estimate that this technique would probably improve the retrieval performance of about 40% of the terms which committed Type-A errors.

II. ERRORS RELATED TO THESAURUS CONSTRUCTION

Most of the indexing errors under investigation which were not attributed to term ambiguity were found to have resulted from imperfections in thesaurus construction. These errors, categorized as Type-B, were due primarily to two sorts of flaws: the inclusion of inappropriate synonymous terms as members of a concept class; and the omission of desirable synonyms for descriptors. These sources of error and a small group of miscellaneous flaws will be examined more closely below.

Errors of False Synonymy

Although all of the index terms within a single concept class, including descriptors, lead-in vocabulary and non-print index terms, are generally referred to as "synonymous terms," there are actually few true synonyms in the "Master's Papers Thesaurus" or, for that matter, in the English language. Aside from spelling variants and, perhaps, technical and common names for the same object or process, few of the words commonly thought of as synonymous are semantically identical. As Dagobert Soergel explains (1974, p. 22) "More often than not, differences in the language indicate subtle differences in meaning or in the affectional component of a concept--for example, Developing countries and Underdeveloped countries." Soergel calls these closely related words "quasi-synonymous terms," and designates them as part of the "equivalence structure," as opposed to the "synonym-homonym structure,"

of the thesaurus.

Some of the members of concept classes in the "Master's Papers Thesaurus" are truly or closely synonymous with the main term, including many which are simply grammatical variants considered necessary for automatic indexing under any but the widest of phrase-match criteria-- e.g. Selection of book for Book selection. However, the index terms allowable in a concept class do not need to be so nearly synonymous to be considered equivalent for indexing and retrieval purposes. Broader term (BT), narrower term (NT), and related term (RT) relations may exist within a concept class just as they do between concept classes in the hierarchy. For example, a component of a precoordinated descriptor may be added as a non-print synonym to improve recall, especially when the component term is unlikely to be used in any unrelated context-- e.g. Mystery for Mystery stories. The synonym is then actually a broader term than the main term.

An NT relationship can occur when a relatively specific term is not considered important enough to merit a separate thesaurus entry. The specific term is then entered as a synonym for a broader descriptor, (e.g. Knitting for Needlework), and is often included as part of the lead-in vocabulary.

A variety of semantic relationships may exist between related terms in a concept class. Related terms may be entered as equivalent terms whenever the distinction between the concepts is not considered important enough to require separate concept classes. This judgement is heavily dependent upon the nature of the document collection and the needs of the user group. For example, in the "Master's Papers Thesaurus," Adult new reader and Illiterates are related terms

belonging to the same concept class. In a thesaurus oriented towards teachers of reading, these would probably be assigned to separate concept classes.

The fact that an index term within a concept class is not actually synonymous with the descriptor representing that class, then, does not necessarily indicate a flaw in the equivalence structure of the thesaurus. Neither, however, does the fact that a term has been designated as an equivalent term guarantee that its meaning will be congruent with the thesaurus-defined meaning of the concept class as a whole.

If the inappropriate usages of a term are occasional and unsystematic, the errors committed should be attributed to term ambiguity. However, some terms designated as equivalent are used in the document collection in senses contrary to the meaning of the concept class as a whole more often than they are used in accordance with the thesaurus-defined meaning. These terms would occasion fewer errors if placed elsewhere in the hierarchy, and therefore were considered flaws of "false synonymy" in the construction of the "Master's Papers Thesaurus." Ten such terms in nine concept classes were responsible for a total of 34 indexing errors. They are listed in the context of each usage in appendix G, table 18.

These false synonyms show all of the relationships to their concept classes mentioned above. The first five listed are actually broader terms to their associated descriptors. Information service and Library science show the standard BT relationship; they encompass and go beyond the concepts of reference service and library science respectively, and could conceivably be hierarchical ancestors of the

other terms in their concept classes.

Environment and Environmental demonstrate a different sort of broader term relationship, being single-word components of a multi-word descriptor, Environmental science. The potential application of these terms is wider than that of the descriptor itself, but the concept represented by the word 'environment' cannot be said to encompass environmental science as 'information service' does reference service.

Reviewers, included as an index term equivalent to Book reviewing and Book reviews, shows two relations at once. It is related to the concepts of reviewing and reviews, and broader than the idea of book reviewing specifically. The BT aspect of the relationship was the source of error here—two documents referred to record reviewers, occasioning four indexing errors.

Training methods and Interviewed by telephone are related to their concept classes in that there is some semantic overlap, but this is neither a generic-specific nor a whole-part relationship. Not all training methods involve staff, for example, and staff training does not always imply methodology.

Two other terms demonstrating RT relationships to their concept classes are single adjectives, Religious and Administrative. These were actually intended to index a number of quasi-synonyms for their descriptors--'religious groups', 'religious organization', etc., for Churches, and 'administrative office', 'administrative aid', etc., for Library director. The actual usage indicates that these words would have been better placed with their cognates, Religion and Administration.

Diary, the sole instance of a false synonym demonstrating a

narrower term relationship to its descriptor, could almost have been tagged an error relating to term ambiguity. A diary is certainly a kind of manuscript, but the implication of an historical document of some kind is missing in its sole textual usage.

Table 18 seems to indicate that equivalent terms broader than the descriptor in a concept class are more likely to occasion indexing errors than NT equivalent terms. To determine, however, whether BT relations are simply predominant in the thesaurus, the index terms in the random sample were examined for the types of relations they displayed. These terms in context for each of their textual usages, are presented in appendix G, table 19.

Forty-seven of the sample terms were descriptors themselves. Of the 53 non-descriptors, 25 were truly or nearly synonymous with their main terms. These entailed 10 spelling variations, including abbreviations and initialisms; 3 syntactical variants (e.g. School of library science for Library science schools); and 12 terms very similar semantically to the descriptor (e.g. Preschooler for Preschool children).

Common sense would indicate that true synonyms and close quasi-synonyms would be unlikely to occasion the type of indexing error under consideration in this section, and this is supported by the evidence of the sample. Nearly half of the equivalent terms in the sample fit this description, indicating a large number of such terms in the thesaurus, yet none of these were involved in errors related to false synonymy.

Thirteen of the index terms in the sample were narrower than their associated main terms. Seven of these were involved in a generic-specific relationship like that between Diaries and Manuscripts.

Two index terms in the sample designated the person involved in the activity represented by the descriptor (e.g. Bookdealer for Book selling). While these may appear to hold an RT relation, the semantic factoring of these concepts reveals that the descriptor is actually the broader term. The basic rule of hierarchy formulation from the combination of concepts holds that a compound concept is narrower than each of its components (Soergel 1974, ch. C). While Book selling factors into the concepts of books and of selling, Bookdealer factors into books, and selling (or 'bookselling',) and person.

The last four NT equivalent terms were component terms of compound descriptors, as in the case of Maps in the concept class designated by Maps and atlases. These compound descriptors should not be confused with compound concepts such as 'book dealers' above, and the 'and' in the descriptor is actually equivalent to the Boolean OR. The rule of hierarchy formulation, of course, applies only to compound concepts in which all semantic factors are required components.

Only five of the equivalent terms in the sample were broader than the descriptor for their concept class. Four of these showed the sort of relationship demonstrated by Reference service and Information service; the descriptor represents a particular type or aspect of the concept designated by the synonym. Three of these terms were included in the lead-in vocabulary, in order to indicate the scope of the concept class to the user. This was unnecessary in the case of the fourth term, Employment, which would have appeared adjacent to the descriptor, Employment practices, in the alphabetical index.

The fifth term, Administrating, was included as a synonym for Library director, which can be factored into three sets of semantic

components: administrating, libraries, and persons. This "synonym" seems especially likely to occasion errors in indexing. Happily, 'administering' is a more common word for the same concept, and the term might never index at all.

These results do reinforce our tentative conclusion that equivalent terms which are broader than their associated main term are more likely to occasion errors of false synonymy. These accounted for exactly half of the terms in table 18, while representing only 9.4% of the 53 non-descriptors in the sample. Equivalent terms which are narrower, or more specific, than their descriptors appear to be relatively safe. Nearly 23% of the 53 sample terms demonstrated this relationship, while only one such term in the thesaurus occasioned this type of error.

The remaining 11 index terms in the sample held RT relationships to the other terms in their concept classes. These included four simple grammatical variants (e.g. Administer for Administration) which might have been classed as quasi-synonyms. They were considered related terms because in some instances the thesaurus maintained separate concept classes for different grammatical forms; it would be improbable for such a distinction between actual synonyms to have any use.

The final seven terms demonstrated the same sort of semantic overlap as Training methods and Staff training. When these concepts are subjected to semantic factorization, it can be seen that one component or set of components is held in common with the descriptor, while the other component(s) of the descriptor and the equivalent term are quite different.

The inclusion of this type of equivalent term effectively widens the scope of the concept class as a whole. With the addition of Hearing impaired as a synonym for Deaf, for example, the descriptor could quite appropriately have been renamed 'Deaf and hearing impaired', if this were not rather useless for the purpose of automatic indexing. The same effect can be achieved more practically when the related term is included as an entry in the alphabetical index. Even so, these related terms may still occasion indexing errors when they have potential applications outside of the limitations imposed on their meanings by their places in the hierarchy. Training techniques, for example, need not always apply to staff, yet it appears as a descendant of Staffing. This illustrates the point that non-descriptors should be placed with as much care and consideration as main terms.

Errors Related to the Precision of Indexing

Whenever a concept found in the document collection was not indexed in any manner by the "Master's Papers Thesaurus," this was designated an error of omission, falling outside of the scope of this study. If, on the other hand, a concept was indexed with one descriptor when a semantically closer concept existed in the thesaurus, this was not invariably counted as an error of commission. When, in the judgement of the investigator, the more appropriate descriptor failed to post because of the omission of a desirable synonym, this was judged an error due to thesaurus construction. However, if the indexing actually performed was judged to be the best that could be accomplished by a simple automatic term-matching procedure with any thesaurus, no error was designated.

While an attempt was made to reach these decisions consistently and systematically, not all judgements could be achieved through the simple application of relatively objective criteria. If, for example, the most appropriate descriptor failed to post because an alternate spelling for the term was not included in the thesaurus, this would certainly constitute a Type-B error. Similarly, if the more appropriate concept were only implied, and associated words appeared nowhere in the document abstract, the thesaurus clearly could not be held culpable.

In actuality, however, all of the cases considered fell somewhere between these two extremes, and most often the decision was not whether a synonymous term could have been included, but whether it should have been. Factors which had to be considered include the practical limitations on the number of equivalent terms in a concept class, the probability of any particular natural-language phrase occurring with any frequency in the document collection, and the accessibility of the specific concept through the post-coordination of existing index terms.

A look at the textual passages indexed by the descriptor Children helps to illustrate the complexity of the problem. Children posted a total of 42 times in the indexing of the 149 master's papers. (This figure does not include one input error, 'childern', which was not indexed at all.) The concept class designated by Children also includes the index terms Young people and Juvenile. Of the textual usages which occasioned the postings of Children, 17 referred to children's literature; four to library services to children; seven to exceptional children; and one to services to exceptional children.

Concept classes for each of these specific topics are included in the thesaurus, as shown in table 10. This table also reproduces the 29 passages, indexed by Children, referring to these more specific ideas. The problem is to determine which of these postings would be called errors of precision due to the omission of a desirable synonym in the more appropriate concept class in the thesaurus.

Some judgements were relatively simple. 'Literature for children', which occurred four times in the document collection, seems to be an appropriate synonymous term for Children's literature, and its omission was counted as a flaw in the construction of the "Master's Papers Thesaurus."

'Children's book' would appear to be another obvious candidate for the concept class designated by Children's literature. However, all of the terms in this class considered together clearly imply that the class relates only to children's fiction, and not to texts or other non-fiction works which might have been designated by 'children's books' in the document collection. The decision in this case was that this distinction was not necessary in the thesaurus, and seven indexing errors were declared.

'Children's picture storybooks' was used twice in the document collection, and might appear to be another acceptable equivalent term for Children's literature. However, both usages occurred in a single document, and it does not seem to be a phrase with much probability of occurring again in the collection. In this case Children was considered adequate, and no errors were declared.

The underlined words in table 10 indicate which phrases were finally judged to constitute improperly omitted synonyms. Passages

TABLE 10.--Determination of errors involving the descriptor Children

Appropriate concept class	Context of document usages	Designated errors
Children's literature	<u>literature for children</u> (4x)	11
Children's writers	<u>children's book field</u>	
Children's fiction	<u>children's books</u> (6x)	
Children's authors	children's picture story-books (2x)	
Juvenile literature	works of fantasy written by [3 author's names] for children	
Juvenile fiction		
Stories	folk and fairy tales for children	
	children's editions	
Exceptional children	<u>visually handicapped children</u> (3x)	5
	<u>physically handicapped children</u>	
	<u>disabled children</u>	
	<u>mentally retarded children</u>	
	recreational reading for children	
Services to children	<u>public library programs for children</u> (3x)	3
Children's service		
Services to young people	children's public library service	
Services to exceptional children	library services as they affect partially seeing children	0
	Total errors	19

with no underlined words were not associated with any errors related to the precision of indexing. Hopefully, this example serves to illustrate the manner in which these decisions were approached.

Table 11 presents the errors finally attributed to flaws in the constitution of the "Master's Papers Thesaurus." The term which actually posted is listed first, followed by the passage of text which occasioned the indexing. As in table 10, underlined words represent phrases considered suitable synonyms for the appropriate descriptor. The descriptor for the preferred concept class and the number of errors involved follow. Errors involving the index term Children are not reiterated here, but are included in the total.

The sum of 45 errors arrived at here is a conservative figure, for the original thesaurus construction was given the benefit of every doubt. However, these errors are not very serious, as they can be remedied by the simple addition of index terms in the next routinely scheduled update. There does not appear to be any particular method of avoiding these flaws in the thesaurus construction stage, other than the careful examination of KWIC index or other representation of a sample of the document collection. Common sense does indicate that it is better to wait for problems to arise based on actual term usage than to try to include every conceivable equivalent term within each concept class.

Miscellaneous Errors

Eleven errors were attributed to flaws in the "Master's Papers Thesaurus" which involved neither false synonymy nor the omission of index terms.

Two descriptors were placed in inappropriate places in the

TABLE 11.--Errors relating to precision of indexing

Posted description	Context of usage(s)	Preferred description	Errors
Adults	<u>new literate adults</u> <u>functionally</u> <u>illiterate adults</u>	Adult new readers	2
Catalogs	<u>distributor's</u> <u>catalogs</u> <u>publisher's</u> <u>catalogs</u>	Bookdealers catalogs	2
	<u>on-line catalog</u>	Automated catalogs	1
Children	(see table 6)		19
Education	<u>education in</u> <u>librarianship</u>	Library education	1
Fiction	<u>historical fiction</u> (2x)	Historical novels	2
Institutes and workshops	<u>technical insti-</u> <u>tutes</u>	Junior colleges and technical schools	1
Literature ^a	<u>scientific lit-</u> <u>erature</u>	Science	2
Pictures	<u>sixteen millimeter</u> <u>picture projec-</u> <u>tors</u>	Equipment and supplies*	1
Publications	<u>place of publica-</u> <u>tion</u> <u>date of publica-</u> <u>tion</u>	Imprints	2
Reading	<u>microform reading</u> <u>room</u>	Facilities*	3
Science	<u>arts and sciences</u>	Liberal arts	1
Selection	<u>selection criteria</u> <u>criteria for</u> <u>selection</u>	Selection policies	4
State libraries	<u>State Library</u> <u>Commission</u>	Governing authorities of libraries	1

TABLE 11.--Continued

Posted description	Context of usage(s)	Preferred description	Errors
Young adults	<u>young adult services</u>	Services to young adults	1
Total errors			45

NOTE: Descriptors marked with an asterisk (*) are slightly too broad to represent entirely appropriate concept classes. The associated errors might be considered errors of omission, due to the underlined phrase not appearing as a thesaurus descriptor.

^aThe other possible precoordinated forms of Literature are not listed here, as each of these except 'scientific literature' occasioned the posting of the subject or field as well as of Literature. Science, due to the lack of synonyms, represents the only inaccessible concept in this group.

thesaurus hierarchy due to an apparent misunderstanding of their scope. Levels of service, placed under Staffing in the "Organization and administration" division of the hierarchy, was used in a document to refer to measures of public service, evidently a more common application. This phrase does not appear in any dictionary or glossary of library science or management terms, and illustrates the need for consulting subject specialists in the construction process. Continuing education, a descendant of Library education in the thesaurus, clearly pertains to non-librarians as well, although this did not occur to the construction group at the time. These descriptors were involved in three errors.

Seven errors were occasioned by descriptors posting to their antonyms. Fiction matched 'non-fiction' five times in two documents, and Periodical matched 'non-periodical' once. This sort of error could be modified to respect hyphens instead of stripping them before beginning the matching process. This might, however, result in the failure of other concepts to index when such posting is desired. Alternatively, 'non-Y' could be added to the concept class of any appropriate index term Y as a posting-prohibiting synonym.

The seventh error of this type admits of no simple solution, and by the criteria employed for determining errors of precision, would not be attributed to a flaw in the thesaurus. A passage of text reading, ". . . not meant as a guide to selection" occasioned the posting of the descriptor Selection. This sort of construction ought to be advised against in guidelines for the writing of abstracts.

The final Type-B error found in the indexing of the master's papers was the posting of a descriptor, Wake County Public Library. This might be termed an empirical mistake, for there is no Wake County

Public Library although a Wake County Public Library System exists.

III. ERRORS RELATED TO THE MATCHING PROCESS

False Term Combinations

Five errors were found to have occurred when multi-word index terms matched unanticipated word combinations in the documents. As so few errors were involved, the circumstances of each are described below:

1. Book number, a descriptor synonymous with Cutter number, posted to a document which contained the passage, ". . . and the number of books in the collection which were written by women."
2. Library and society, posted to a document entitled "The Society Hill Library Society of South Carolina."
3. Library laws, a non-print synonym for Library legislation, indexed a document which contained the following list: "1) inventories of attorney's libraries, (2) a law book catalog, (3) citations from the reports of Maryland, [etc.]" The underlined passage triggered the indexing.
4. User surveys indexed a passage which read, "[This book] is recommended and presented as a means of bridging the gap between media resources and users. A survey of recent literature establishes . . ." Again, the responsible words are underlined.
5. AMERICAN LIBRARIES, the proper name of a publication, indexed the passage, "the American library profession." This resembles the Type-B errors relating to the precision of indexing, as it appears that adding the synonym 'library profession' to the concept class

Professionalism would remedy the situation. Because of the order in which term matches are sought in the indexing system, however, AMERICAN LIBRARIES will always index first.

None of the first four errors would have occurred under the strictest of the phrase match criteria available in the Automatic Indexing system (see appendix B). However, all of these errors except the first would have been committed under the next strictest option, which specifies that index term words must appear in the document in order and adjacent to each other except for the interposition of deleted stop words. (In the case of the third error (3) only the stop word 'a' actually comes between 'library' and 'law', for punctuation marks and numbers are deleted prior to matching.)

The first error listed, involving Book number, could only have occurred under the two most liberal options for matching phrases available in the system. One of these, the option actually used for the indexing of the master's papers collection, specifies that words must be adjacent, but allows them to appear in any order and disregards intervening stop words. While performing indexing with a stricter phrase match criterion would have prevented this false match, a single error seems a small price to pay for the added flexibility granted by the less restrictive option.

The fifth error, involving AMERICAN LIBRARIES, could be prevented from recurring by the addition of an index term, 'American library profession', to the most appropriate concept class, Professionalism. However, this omission would not have been classed as an error by the criteria given in "Errors Relating to the Precision of Indexing" above. It is included here as an instance of a second sort of failing

to which the automatic phrase match procedure is susceptible.

These five errors represent just over 2% of the total number of errors of commission discovered in the indexing of the document collection. The fact that so few errors of this nature were committed, even under the relatively lenient stipulations of the phrase match criterion employed, indicates that unanticipated word combinations are not a major problem in automatic indexing.

Input Errors

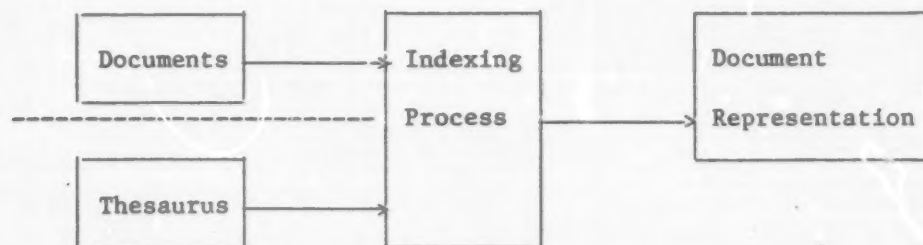
Mistakes in the input texts of the documents were responsible for two final errors. While several misspellings and input errors were found in the course of this study, most of these simply prevented term matches from taking place, and so were responsible for errors of omission only. In two instances, however, the input error involved only one word in a precoordinated descriptor the other word of which was also a discrete index term. In one case, 'Selection aides' in the text was indexed by Selection instead of Selection aids, and in the other, 'anntated bibliography' was indexed by Bibliography rather than Annotated bibliography.

These actually constitute errors related to the precision of indexing as discussed above. However, they must be attributed to faulty data, and not to any flaw in the "Master's Papers Thesaurus."

IV. CONCLUSION

This study has been concerned with a single aspect of thesaurus-based automatic indexing--the coincidence of the natural language of a document collection and the controlled vocabulary of a thesaurus. The indexing system has been viewed as a simple process of the convergence of language, as illustrated in figure 3. As input, we have the text, a natural language, and the thesaurus, an indexing language; as output, we get the document representation, a sort of shorthand for the two. If the representation is faithful to both the thesaurus and the document, the indexing is correct; otherwise indexing failure has occurred.

Natural language



Indexing language

Fig. 3. The automatic indexing system

In this investigation 230 indexing errors were identified. These errors were classified into three broad categories according to the underlying reasons for failure, which correspond roughly to the components of the indexing system: term ambiguity (the natural

language), thesaurus construction (the indexing language), and the mechanics of the term matching procedure (the indexing process).

Within some of these categories of error, sub-categories were identified, also in accordance with reasons for failure. Errors related to thesaurus flaws were found to occur mainly because of the inclusion of inappropriate synonyms in concept classes, and the omission of desirable synonyms from concept classes. Errors related to the matching procedure were found to arise from imperfections in input and semantically unrelated term combinations in the texts. Attempts to categorize errors related to term ambiguity in this manner were less satisfactory. Reasons for errors of ambiguity, sought in either the nature of the ambiguity displayed or the technicality of the vocabulary, did not correlate well with observed patterns of failure. In the end, Type-A errors had to be discussed in terms of what could be done to prevent them, resulting in a somewhat slippery categorization, with some errors falling into more than one class.

The distribution of errors arrived at in this manner is summarized in table 12. In general, this typology presented a useful framework for the analysis of errors, allowing particular problems, such as including terms hierarchically broader than their descriptors in concept classes, to be pointed out. Solutions to problems, such as employing posting-prohibiting terms to prevent undesirable indexing, could also be noted.

It should be emphasized, however, that alternate approaches to indexing failure could have been employed, leading to different results. For example, this study takes indexing errors as the basic unit for classification. The index terms involved in errors are another

TABLE 12.—Distribution of indexing errors among major categories and sub-categories

Category Totals		Type of Error	Subcategory Totals		
Number of errors	Percentage of all errors	CATEGORY Subcategory	Number of errors	Percentage of category errors	Percentage of all errors
133	57.8	TYPE-A	. . .	. . .	. . .
90	39.1	TYPE-B			
		False synonymy	34	37.8	14.7
		Precision	45	50.0	19.6
		Miscellaneous	11	12.2	4.8
7	3.1	TYPE-C			
		Term combination	5	71.4	2.2
		Input	2	28.6	0.9

possible starting point. Location in the thesaurus hierarchy, with reference to either the general subject divisions or the levels of descent (representing the specificity of terms) might show a suggestive correlation with patterns of failure.

In addition, the scope of this investigation is limited to a subset of all possible indexing failures, designated here as "errors of commission." A lack of congruence between the indexing language and the language of the document collection, however, can occasion other types of errors as well. For example, a synonym for a descriptor

may be omitted, with the result that a desired concept is not indexed at all.

Furthermore; the indexing system actually encompasses far more than figure 3 indicates. This study does not challenge the conceptual adequacy of the thesaurus in relation to genuine user needs. It has been assumed that the indexing was "correct" whenever the meaning of a descriptor matched the meaning of the passage responsible for its posting. Whether this descriptor truly represented the content of the document, or attained the levels of specificity or exhaustivity required by the typical user, was not questioned.

The conceptual organization of the thesaurus can occasion indexing failures of another nature as well. Imperfections in the hierarchical arrangement of concept classes can cause problems in mechanized retrieval, when a hierarchically broader term is "exploded," or expanded to include all of the more specific terms associated with it. Even if the more specific descriptor perfectly reflects the meaning of the document it represents, the document may be unrelated to the more generic descriptor.

Finally, even if the document representation achieved through the indexing is technically and conceptually perfect, the quality of retrieval is not guaranteed. Not all index terms are equally helpful in distinguishing between relevant and non-relevant materials. In retrieval systems which operate by calculating measures of correlation between the index terms assigned to documents and those assigned to requests, index terms with certain patterns of distribution in the collection may actually lower retrieval performance (Salton 1975).

This study, then, considers only one aspect of indexing

failure, and that from a particular point of view. Other approaches may be equally productive, and are certainly necessary for a comprehensive understanding of automatic thesaurus-based indexing. Hopefully, however, the results and conclusions presented here will be of aid to those involved in the design and evaluation of indexing systems.

APPENDIX A

OUTLINE OF THE HIERARCHICAL STRUCTURE OF THE "MASTER'S PAPERS THESAURUS"

Underlined terms are major division labels. The index terms listed within each division are superior nodes of all main hierarchical branches.

General

- Censorship
- Communications
- Copyright
- Library and society
- Library education
- Library history
- Library legislation
- National organizations
- Professionalism
- Types of libraries
- Types of literature
- Types of users

Bookmaking, publishing and the book trade

- Bookmaking
- Book trade

Library materials

- Books
- Serials
- Series
- Proceedings
- Dissertations
- State-of-the-art
- Technical reports
- Photocopies
- Non-book materials
- Special collections

Personal name authority list

Schools

Organization and administration

- Administration
- Governing authorities of libraries
- Budget and finance
- Library committees
- Staffing
- Staff training
- Equipment and supplies
- Library statistics
- Personnel
- Public policy
- Building and grounds

Public services

- Circulation
- Interlibrary loan
- Library promotion
- Library services
- Reference service
- Shelving
- User surveys

Technical services

- Bibliographic control
- Catalog maintenance
- Collection management

Geographical list

Chronology

Terms of general application

APPENDIX B

PHRASE MATCH CRITERIA

The following phrase match criteria, listed in order from least to most restrictive, are matching options available in the Automatic Indexing system.* Thesaurus index terms are underlined; stop words are enclosed in parentheses.

1. The words in a multiple-word index term may appear anywhere in a document text.

Example: Library society and
Library (and) society both match
"(the) literary society met every
week (in) (the) public library"

2. The words in an index term may appear in any order in the text, but must be adjacent except for intervening stop words.

Example: Library society and
Library (and) society both match
"(the) library (in) society" or
"(the) society (and) (the) library"

3. The words in an index term may appear in any order in the text, but must be adjacent in the text. Stop words in an index term will match any word in the text which meets the adjacency requirement.

Example: Library society matches
"library society" or
"society library" only

Library (and) society matches
"library [any word] society" or
"society [any word] library" or
"[any word] library society" etc.

4. The words in an index term must appear in order and adjacent to each other in the text, except for intervening stop words.

Example: Library society and
Library (and) society both match
 "library society" or
 "library (and) society" or
 "library (in) (the) society" etc.

5. The words in an index term must appear in order and adjacent in the text. Stop words in the index term will match any word occupying the same position in the text.

Example: Library society matches
 "library society" only

Library (and) society matches
 "library [any word] society" only; not
 "library society"

*Description adapted from the "User's Manual for the Automatic Indexing System" (Strode 1976, appendix A).

APPENDIX C

RANDOM SAMPLE OF 100 INDEX TERMS IN THE "MASTER'S PAPERS THESAURUS"

Sample terms are listed in the order in which they appear in the thesaurus. For each sample term, the following information is given:

- col. 2 - Type of index term. Abbreviations for type of term are as follows:

des = descriptor
lea = lead-in term
npt = non-print term

- col. 3 - Descriptor for concept class of which the sample term is a member, if the sample term is not itself a descriptor

- col. 4 - Number of matches found by sample term in the document collection

- col. 5 - A check mark indicates that sample term is a "technical" library science term, included as an entry in the British Library Association's Librarian's Glossary or the ALA Glossary of Library Terms. The symbol appears in parenthesis (✓) if the glossary entry term is a variant form of the index term.

Sample term	Type of Term	Descriptor (if <u>lea</u> or <u>npt</u>)	Number of Matches	Technical Term
Confidentiality	des	. . .	0	
Sexism	des	. . .	0	
Cable television	des	. . .	0	
Radio	des	. . .	0	
Signs	des	. . .	0	
Copyright	des	. . .	1	✓
Copyright problems	des	. . .	1	
Copyright fee structure	des	. . .	0	(✓)
Library Services Act	des	. . .	2	✓
AMERICAN LIBRARIES	des	. . .	2	
Unionization	des	. . .	0	
University libraries	lea	Academic libraries	3	✓
Robert Burton House	des	. . .	0	
Undergraduate Library				
Robert Burton House Library	npt	Robert Burton House Undergraduate library	0	
Academic affairs libraries	des	. . .	1	
Architecture	des	. . .	0	
Crafts	des	. . .	1	
Novels	lea	Fiction	5	✓
Biographical	npt	Biography	1	(✓)
Business administration	des	. . .	0	
Mathematical	npt	Mathematics	0	
Reluctant reader	des	. . .	2	
Pre-schooler	npt	Preschool children	0	
North American Indians	npt	Indians	0	
Institutions	des	. . .	0	
Hearing impaired	lea	Deaf	0	
Printing	des	. . .	12	✓
Illumination	npt	Book illumination	0	✓
Book marketing	npt	Book marketing and promotion	0	
Book prices	des	. . .	0	
Book fairs	des	. . .	0	✓
Book dealer	lea	Book selling	0	
Enumerative bibliography	des	. . .	0	✓
Dictionary	des	. . .	1	✓
Handbook	des	. . .	0	✓
Thesaurus	des	. . .	0	✓
Theses	lea	Dissertations	1	✓
Filmstrips	des	. . .	3	✓
Video tapes	npt	Videotapes	0	(✓)

Sample Term	Type of Term	Descriptor (if <u>lea</u> or <u>npt</u>)	Number of Matches	Technical Term
DIALOG	des	. . .	6	
Manuscripts	des	. . .	8	✓
Diaries	lea	Manuscripts	1	
Map	npt	Maps and atlases	0	✓
United States government publication	npt	Federal publications	0	
Census report	npt	Bureau of the Census publications	0	
International documents	npt	International publications	0	
UN document	npt	United Nations publications	0	
Organization and administration	des	. . .	1	
Administer	npt	Administration	0	
Library boards	lea	Governing boards	1	✓
Fund drive	des	. . .	1	
Levels of service	des	. . .	1	
Employment	npt	Employment practices	3	
Supervisor	npt	Supervision	1	✓
Evaluation of personnel	npt	Staff evaluation	0	
Evaluating personnel	npt	Staff evaluation	0	
Personnel evaluation	lea	Staff evaluation	0	
Longevity	npt	Length of service	0	
Training of staff	npt	Staff training	0	
Training techniques	npt	Staff training	0	
Training management	npt	Leadership training	0	
Administrating	npt	Library director	0	
Public librarians	des	. . .	2	
Public relations	des	. . .	5	
Library building	npt	Library construction	3	
Systems of security	npt	Security systems	0	

Sample Term	Type of Term	Descriptor (if <u>lea</u> or <u>npt</u>)	Number of Matches	Technical Term
Library services	des	. . .	18	✓
Program administration	des	. . .	0	
Services to adults	des	. . .	1	
Services to business	des	. . .	0	
Tours of the library	npt	Library	1	
I & R	npt	orientation Information and referral services	0	(✓)
Information storage and retrieval	lea	Information retrieval system	1	
Mechanical abstracting	npt	Automated abstracting	0	
LC Subject Heading	npt	Library of Congress Subject Heading	0	
Sear's	npt	Sears List of Subject Headings	0	
Cataloging convention	npt	Cataloging rules	1	
International Standard Book Number	lea	ISBN	0	✓
Centralized cataloging	lea	Cooperative cataloging	0	✓
Card catalogue	npt	Card catalogs	1	✓
Subject catalogs	des	. . .	0	✓
Book number	des	. . .	1	✓
Faceted class	npt	Faceted clas- sification	0	✓
Approval plan	npt	Approval and gathering plans	0	
Want lists	lea	Desiderata	0	✓
Africa	des	. . .	2	
Illinois	des	. . .	2	
S.C.	npt	South Caro- lina	0	
Colonial	des	. . .	7	
Preschool	npt	Preschool and day care	0	
Secondary schools	des	. . .	5	
Two-year colleges	lea	Junior col- leges and technical schools	3	

Sample Term	Type of Term	Descriptor (if <u>lea</u> or <u>npt</u>)	Number of Matches	Technical Term
Duke University	des	. . .	3	
UNC-Chapel Hill	npt	University of North Caro- lina	1	
Computer programs	des	. . .	0	
Computer hardware	des	. . .	0	
LeGuin, Ursula	des	. . .	1	
School of library science	npt	Library sci- ence schools	4	(✓)
Educational	npt	Education	6	
Children's library	lea	Elementary school libraries	0	

APPENDIX D

CONTEXT OF INDEX TERMS INVOLVED IN TYPE-A ERRORS

Table 13 lists all of the index terms involved in errors related to term ambiguity. The context of each textual usage to the first stop word or punctuation mark is shown. Where this is not enough to indicate the meaning of the usage, additional text is given.

Term occurrences involving Type-B and Type-C errors are not included in this table. Correctly indexed extracts are indicated first, followed by a row of stars and the passages occasioning Type-A errors. The term itself is represented by a ~.

Possible precoordinated forms as noted in "Precoordination to Resolve Ambiguity," are indicated in the final column.

TABLE 13.--Terms involved in Type-A errors--context of all usages
occasioning no errors or Type-A errors

Index term	Document usage	Possible precoor- dinated form
Acquisition*	given to the ~ and care selection and ~ . behind their ~ . microform ~ include except that the ~ of fiction . operation of ~ . pattern of ~ and growth area of ~ . ~ were divided ***** librarian in the ~ of library	aim to
Aim*	changing ~ and objectives ***** position and ~ to continually books which ~ to describe	
AMERICAN LIBRAR*	--- ***** documents in ~	
Bibliograph*	indexes and ~ . authoritative ~ and book review dissertation ~ and indexes multitude of ~ and producer's various ~ and the major directories and ~ could be used include ~ of the books catalogs and ~ . cross-references and ~ . dissertation ~ . precedes the ~ . index, ~ , This ~ comprises education ~ identified The ~ includes this ~ provides citations in the ~ were located audiovisual ~	

TABLE 13.--Continued

Index term	Document usage	Possible precoor- dinated form
Catalog*	non-book ~ . ***** European ~ in general. development of ~ . CURRENT CATALOG mainly on ~ cards chronological ~ of the North manuscript ~ and bibliographies INDEX-CATALOGUE, in the ~ were changed ***** libraries ~ and classify school ~ . ~ of the school (2x)	personnel classifi- cation
Classification*	The ~ of government documents The ~ of United States ***** for that ~ . Factor Analysis ~ and its uses Factor Analysis ~ . personnel ~ which is currently --- ***** indicate that a ~ of change cultural ~ of the United States	
Contemporar*	portray ~ American parents in ~ and Depression ~ bibliotherapeutic indicate that ~ father ~ parents twenty from the ~ period books of ~ poetry review of ~ poetry instiumental ~ music examples of ~ instrumental ***** located in the ~ literature	

TABLE 13.--Continued

Index term	Document usage	Possible precoordi- nated form
Dictionary*	--- ***** authority file, or ~ ,	library fines
Economic*	--- ***** ~ and political and ~ characteristics social and ~ pressures The socio-economic upheavals	
Fine*	library ~ , library ~ : policy such as ~ needs ***** dealers in ~ and rare	
Grant*	State ~ . ***** Walker-Grant Middle School (2x)	
Index*	names ~ was chosen names ~ and a personal each ~ and reviewing each ~ , The INDEX-CATALOGUE, conventions of ~ preparation There is an ~ to the names 19 ~ and 17 printed book catalogs concurrent ~ , regular ~ and bibliographies dissertation ~ are discussed citation ~ for the literature citation ~ offer a unique bibliographies and ~ . ~ , review media, ***** obtains, ~ and disseminates	
Journal*	and ~ articles core ~ collection	

TABLE 13.--Continued

Index term	Document usage	Possible precoor- dinated form
Leadership	in ~ collection needs and ~ usage and ~ . lists of ~ to be added list of ~ to be considered and ~ to be removed ***** ships' logs and ~ , lack of clear ~ and objectives provide strong ~ in attempting Under the ~ of its first consistent ~ on the part public library ~ , ***** position of ~ in service Durham's ~ in library	the liter- ature
Library building*	development of the ~ project ***** present ~ . 1933 a ~ was constructed	
LIBRARY LITERATURE	indexed in ~ , ***** reflected in the ~ . evidenced in ~ . examples from ~ . described in ~ .	
Literature*	article on ~ . geography in ~ for the elementary using ~ to enrich ***** the ~ . . . (26x) science ~ social science ~ library science ~ (3x) environmental science ~ medical ~ (4x) humanities ~	

TABLE 13.--Continued

Index term	Document usage	Possible precoordinated form
	periodical ~ (3x) crocheting ~ contemporary ~ , indexed ~ are examined instruction ~ are found existing ~ types of ~ available of recent ~ establishes in such ~ . extensive ~ on MBO recent ~ on the subject dissertation ~ examines ~ relating history and ~ reveals ~ searching sexually oriented ~ . professional ~ . technical ~ .	
Objective*	~ , and criteria ~ , and reasons ~ , and conclusions emphasizing its ~ and goals goals and ~ . aims and ~ . aims and ~ of the selection demonstrate their ~ and limitations program ~ and to the amount educational ~ for library school leadership and ~ . ***** systematic and ~ collection development	
Picture*	----- ***** national ~ and North Carolina accuracy of this ~ and to provide library ~ in Virginia broad ~ of the role	

TABLE 13.—Continued

Index term	Document usage	Possible precoor- dinated form
Philosoph*	----- ***** programming ~ for young adults	
Publication*	place of ~ , religious ~ (4x) value of such ~ is strongly specific ~ ; reviewing ~ ; use of these ~ can be facilitated percentage of ~ could be able part of the ~ in the field staff and the ~ of the library "scientific" ~ , retrieving ~ that have referred ***** to cease ~ of AIR POLLUTION ABSTRACTS service and the ~ of AIR POLLUTION stimulated the ~ of law	
Reader*	young ~ (3x) causes of ~ failure prevent the ~ from locating ~ frustration fault of the ~ . recommended for the ~ of North ***** READER IN LIBRARY ADMINISTRATION (2x) reveal to the ~ general acquaints the ~ with the	
Reading*	and ~ , derision for ~ a book effects of ~ fiction significance of ~ fiction public in their non-reading hours patterns of ~ . meeting the ~ needs teaching ~ ***** recreational ~ (2x)	

TABLE 13.—Continued

Index term	Document usage	Possible precoordinated form
Record*	using ~ materials popular ~ matter psychicworks, or ~ , psychic ~ . SCHWANN RECORD AND TAPE GUIDE ~ reviewing (4x) aids to ~ selection THE NEW RECORDS ***** presidential ~ (3x) reference question ~ (3x) ~ appear to be used made of the ~ by the individuals LC ~ found standardize the ~ in the OCLC data extant ~ is scarce keep ~ of reference questions Colonial ~ Project catalogue ~ used	
Selection*	problem(s) of ~ (3x) participating in ~ (2x) record ~ , ~ and use of these publications processes of ~ and weeding guide to the ~ initial ~ by a public library ~ criteria (2x) criteria for ~ (2x) responsibility for ~ (2x) used in the ~ of jazz reviewing for the ~ of recordings ~ of simulation games methods of ~ presently ~ procedures the ~ process (2x) ***** careful ~ of appropriate ~ of a format with the ~ of the most	

TABLE 13.--Continued

Index term	Document usage	Possible precoor- dinated form
Series	Children's ~ books known ~ by author authors and ~ discovered different ~ , books in ~ . ***** findings of a ~ of personal second ~ of retrievals	series of
Student*	within the ~ body ~ , teacher, and librarian status of each ~ . develop a ~ profile STUDENT-TEACHER GUIDE (3x) art ~ use those of the ~ . ~ and library personnel student's attitudes why ~ chose all ~ enrolled elementary school ~ (2x) most ~ indicated indicated that ~ do use available for ~ in grades suitable for ~ in grades ~ most often materials by the ~ of the ~ , either the ~ or the teacher reaction of ~ . and ~ seeking one for ~ , needs of ~ , information on ***** the ~ mail evaluation of a ~ workbook	
Subject analysis	--- ***** comprehensive ~ of the holdings	

TABLE 13.--Continued

Index term	Document usage	Possible precoor- dinated form
Western*	---- ***** Bell Labs and ~ Electric sites	

APPENDIX E

HAZARDOUS TERMS IN THE "MASTER'S PAPERS THESAURUS"

A corpus of hazardous terms was compiled from all of the index terms in the thesaurus which occasioned the posting of at least one descriptor. Thirty of the terms identified in this manner were responsible for Type-A errors, and are listed, with the context of each usage, in appendix D. The 36 hazardous terms involved in no errors are listed below. Terms are given in the singular form only; terms not appearing as dictionary entries are indicated by an asterisk (*).

Abstract	Facility
Accreditation	Fantasy
Administration	Handicapped
Allocation	Humanity
Black	*ILL
Blind	*LIBRARY JOURNAL
Budget	Management
Circulation	Network
*Classification system	Novel
Colonial	Planning
Craft	Press
Depression	Reformation
Development	*SEARS
*DIALOG	*Selection policy
Display	Slide
Employment	Survey
Engineering	Theses
Evaluation	Weeding

Tables 14 and 15 indicate the division of all hazardous terms into homographs and polysemes, and into truly and contextually ambiguous terms, respectively. Only terms appearing as dictionary entries

were classified. These tables correspond to the discussion in
"Classification According to the Nature of Ambiguity."

TABLE 14.--Classification of hazardous terms into formal homographs and polysemes

	Homographs	Polysemes	
Involved in errors	Fine* Grant*	Acquisition* Aim* Bibliograph* Catalog* Classification* Climate* Contemporar* Dictionary* Index* Journal* Leadership* Literature* Objective*	Picture* Philosoph* Publication* Reader* Reading* Record* Ser*(ies) Student* Selection* Western*
Not involved in errors	Novel* Press*	Abstract* Accreditation Administration* Allocation* Blind* Budget* Circulation* Colonial* Craft* Depression* Development* Display* Employment Engineering	Evaluation* Facilit* Fantas* Geograph* Handicapped Humanit* Network* Planning Reformation Slide* Survey* Theses Volunteer* Weeding
Total terms	4	51	
Involved in errors	2	23	
Percentage of total	50.0%	45.1%	
Not involved in errors	2	28	
Percentage of total	50.0%	54.9%	

TABLE 15.--Classification of hazardous terms into truly and contextually ambiguous terms

	First Classification		Second Classification	
	Truly ambiguous	Contextually ambiguous	Truly ambiguous	Contextually ambiguous
Involved in errors	Aim*	Acquisition	Aim*	Acquisition*
	Bibliograph*	Catalog*	Bibliograph*	Catalog*
	Climate*	Classification*	Fine*	Classification*
	Contemporar*	Leadership	Grant*	Climate*
	Dictionar*	Selection*	Index*	Contemporar*
	Fine*		Objective*	Dictionar*
	Grant*		Reader*	Journal*
	Index*		Reading*	Leadership
	Journal*		Western*	Literature*
	Literature*			Philosoph*
	Objective*			Picture*
	Philosoph*			Publication*
	Picture*			Record*
	Publication*			Selection*
	Reader*			Ser*(ies)
	Reading*			Student*
	Record*			
	Ser*(ies)			
	Student*			
	Western*			

TABLE 15.--Continued

	First Classification		Second Classification	
	Truly ambiguous	Contextually ambiguous	Truly ambiguous	Contextually ambiguous
Not involved in errors	Abstract*	Allocation*	Abstract*	Administration*
	Accreditation	Evaluation*	Accreditation	Allocation*
	Administration*	Planning	Blind*	Black*
	Black*		Budget*	Circulation*
	Bind*		Craft*	Colonial*
	Budget*		Engineering	Depression*
	Circulation*		Facilit*	Development*
	Colonial*		Handicapped	Display*
	Craft*		Humanit*	Employment
	Depression*		Novel*	Evaluation*
	Development*		Press*	Fantas*
	Display*		Slide*	Network*
	Employment		Theses	Planning
	Engineering			Reformation
	Facilit*			Survey*
	Fantas*			Volunteer*
	Handicapped			Weeding
	Humanit*			
	Network*			
	Novel*			
	Press*			
	Reformation			
	Slide*			
	Survey*			
	Theses			
	Volunteer*			
	Weeding			

APPENDIX F

PRE- AND POST-COORDINATION OF TERMS

Tables 16 and 17 deal with the relative effectiveness of pre-coordinated index terms and the post-coordination of single-word terms. Both tables exclude index terms which are proper names and all of the index terms consisting of a single content word and one or more stop words. 'Library' is considered a stop word here due to its high frequency of occurrence in the document collection. 'Library' itself is not an index term appearing in the thesaurus only as a component of precoordinated terms.

Table 16 lists the precoordinated index terms in the "Master's Papers Thesaurus" which found at least one match in the document collection. The following information is included:

- col. 1 Index term in the form in which it appears in the thesaurus.
- col. 2 Total number of matches found for the precoordinated term, counting both singular and plural forms.
- col. 3 Precision of the precoordinated term (total correct hits/total matches).
- col. 4 Uniterm components of precoordinated index term, exclusive of stop words and 'library'.
- col. 5 Occurrences in the document collection of uniterm in col. 2, whether indexed or not.
- col. 6 Documents, designated by their assigned identification numbers, in which the terms in col. 4 co-occur. If the uniterms were combined with a Boolean AND, these are the documents which would be retrieved. A

plus (+) indicates that the meaning of the pre-coordinated index term (col. 1) is appropriate to the document; a minus (-) indicates that it is not. This column may be interpreted broadly as indicating relevant and non-relevant documents retrieved in a hypothetical post-coordinate system. However, the limited implications of "relevance" in this sense should be kept in mind.

col. 7 Precision of retrievals indicated in col. 6 (total relevant retrieved/total retrieved).

Table 17 lists all of the precoordinated index terms in the random sample which did not occasion the posting of any descriptor. Information given is comparable to that in table 16, except that total matches and precision for the precoordinated terms are not listed, as these terms found no matches in the document collection.

TABLE 16.--Precoordinated index terms which found matches in the document collection and co-occurrence of uniterm components

1	2	3	4	5	6	7
Index term	Hits	P	Uniterms	Occurrences	Shared documents	P
Academic affairs library	1	1.00	Academic* Affair*	19 2	1029+	1.00
Book number	1	0.00	Book* Number*	143 21	980- 984- 1006- 1019- 1033- 1041-	0.00
Card catalogue (Card cata- logs)	1	1.00	Card* Catalogue* (Catalog)	2 7 26	984- 1050+	0.50
Cataloging convention	1	1.00	Cataloging Convention*	4 2	1036+	1.00
Copyright problems	1	1.00	Copyright* Problem*	2 29	950+	1.00
Fund drive	1	1.00	Fund* Drive*	11 1	997+	1.00
Information storage and retrieval	1	1.00	Information Storage Retrieval*	68 2 11	1043+	1.00
Levels of service	1	0.00	Level* Service*	16 99	977+ 1014- 1016+ 1065-	0.50
Library Services Act	2	1.00	Service* Act*	99 11	983+ 1004- 1060+	0.66
Organization and administration	1	1.00	Organization* Administration*	8 7	1036+	1.00

TABLE 16.--Continued

1	2	3	4	5	6	7
Index term	Hits	P	Uniterns	Occurrences	Shared documents	P
Public librarians	2	1.00	Public Librarian*	126 73	997+ 998- 1000- 1011+ 1015+ 1016+ 1017+ 1022- 1031+ 1041+ 1048+ 1072+	0.58
Public relations	5	1.00	Public Relation*	126 6	941+ 1039+	1.00
Reluctant reader	2	1.00	Reluctant Reader*	2 15	949+	1.00
School of Library Science	4	1.00	School* Science*	64 36	990+ - 991+ 996- 1018- 1034- 1045- 1059+	0.43
Secondary schools	5	1.00	Secondary School*	7 64	954+ 962+	1.00
Services to adults	1	1.00	Service* Adult*	99 28	974- 1014+ -1079+	0.66
Two-year colleges	3	1.00	Two Year* College*	33 27 15	961+ -1077-	0.50
Precision:						
Macro		0.88				0.75
Micro		0.92				0.58

TABLE 17.--Precoordinated index terms in random sample finding no matches in document collection and cooccurrence of uniterm components

1	2	3	4	5
Index term	Uniterms	Occur- rences	Shared docu- ments	Preci- sion
Approval plan	Approval Plan*	0 3	-	-
Book dealers	Book* Dealer*	143 1	972+	1.0
Book fairs	Book* Fair*	143 0	-	-
Book marketing	Book* Marketing	143 0	-	-
Book prices	Book* Price*	143 0	-	-
Business administration	Business* Administration*	2 7	-	-
Cable television	Cable* Television*	0 5	-	-
Census report	Census* Report*	5 32	-	-
Centralized cataloging	Centralized Cataloging	1 4	-	-
Computer hardware	Computer* Hardware	5 0	-	-
Computer programs	Computer* Program*	5 54	-	-
Copyright fee structure	Copyright* Fee* Structure*	2 0 4	-	-
Enumerative bibliography	Enumerative Bibliograph*	0 38	-	-
Evaluating personnel	Evaluating Personnel	2 10	-	-

TABLE 17.--Continued

1	2	3	4	5
Index term	Unitersms	Occur- rences	Shared docu- ments	Preci- sion
Evaluation of personnel	Evaluation*	36	-	-
Personnel evaluation	Personnel	10		
Faceted class	Faceted	1	-	-
	Class*	2		
Hearing impaired	Hearing*	1	-	-
	Impaired	0		
International Standard	International	2	-	-
Book Number	Standard*	20		
	Book*	143		
	Number*	21		
International documents	International	2	-	-
	Document*	34		
LC subject heading	LC	0	-	-
	Subject*	42		
	Heading*	3		
Mechanical abstracting	Mechanical	0	-	-
	Abstracting	5		
Program administration	Program*	54	1039+	1.0
	Administration*	7		
Services to business	Service*	99	958+	1.0
	Business*	2		
Subject catalog	Subject*	42	936-	0.0
	Catalog*	26	972-	
	(Catalogue*)	7	984-	
Systems of security ^a	System*	45	943+	1.0
	Securit*	1		
Training technique	Training	6	-	-
	Technique*	12		
Training management	Training	6	-	-
	Management*	0		

TABLE 17.--Continued

1	2	3	4	5
Index term	Uniterms	Occurrences	Shared documents	Precision
Training of staff ^b	Training Staff	6 9	1068+	1.0
UN document	UN Document*	0 2	-	-
United States government publications	United States Government* Publication*	22 8 24	-	-
Video tapes	Video Tape*	0 3	-	-
Want lists	Want* List*	0 24	-	-
Precision = Macro				.83
Micro				.63

^aSystems of security is a grammatical variant of the descriptor Security systems, which did post to document 943. Furthermore, under the phrase matching option used for this indexing, there is no difference in the way the two index terms are matched. The recall advantage implied in col. 4 is therefore misleading.

^bTraining of staff is a grammatical variant of the descriptor Staff training, which did post to document 1068. This represents the same situation as Systems of security above.

APPENDIX G

SEMANTIC TERM RELATIONS WITHIN CONCEPT CLASSES

Tables 18 and 19 deal with the semantic relationships between index terms in concept classes, as discussed in "Errors of False Synonymy" in Chapter II.

Thirty-four Type-B errors occurred because concept classes contained index terms inappropriate to the thesaurus-defined meanings of the concept classes as a whole. Table 18 lists all index terms responsible for such errors in column 2, followed by the context of each error-occasioning usage. The number of matches found by each false synonym and the number of errors occasioned are indicated in column 4. The relationship noted in the final column is that of the false synonym to the other index terms in the concept class, which are listed in column 1. The first index term shown in column 1 for each false synonym is the descriptor for that concept class.

Table 19 lists every index term in the random sample which is not itself a descriptor, grouped according to the term relations discussed in "Errors of False Synonymy." These terms are followed by their associated descriptors. Other equivalent terms for these descriptors are not listed here, but may be found in the update printout of the "Master's Papers Thesaurus." The concept class identifiers used in the update are noted in the final column of table 18, to facilitate access to this information.

TABLE 18.--Synonym-descriptor relations--Error-causing terms

1 Other terms in concept class	2 False Synonym	3 Context or use	4 Postings- errors	5 Rela- tion
Book reviewing Book reviews	Reviewer*	(re:record reviewing) (4x)	5-4	BT-RT
Environmental science	Environment*	learning ~ high school ~ interactive ~ (re:location)	5-4	BT
	Environmental	(re:conditions of microform use) (3x)	3-3	BT
Library edu- cation Education for librarians	Library science	~ literature (3x) ~ procedures ~ methodology field of ~	6-6	BT
Reference service Inquiry referral service Reader services	Information service*	(as part of NTIS) (2x) (re:Bell Lab's service) community ~	6-4	BT
Churches	Religious	(re:African religion) (3x) (re:literature) ~ publication (4x)	8-8	RT
Library director Administrator Administrating	Administrative	~ factors ~ setup	2-2	RT
Staff training Training of staff Training tech- niques Methods of training Techniques of training Develop staff Staff improve- ment Staff develop- ment	Training method*	(re:library students)	1-1	RT

TABLE 18.--Continued

1 Other terms in concept class	2 False Synonym	3 Context or use	4 Postings- errors	5 Rela- tion
Telephone reference Telephone interview Interview by telephone	Interviewed by telephone	(re:user survey)	1-1	RT
Manuscripts Logs Autographs	Diar*	questionnaires, ~ , interviews, (as evaluative method)	1-1	NT

TABLE 19.—Synonym-descriptor relations--Random sample

1 Relation	2 Synonym	3 Descriptor	4 Concept class
TRUE AND QUASI-SYNONYMS			
Spelling variations	Card catalogue	Card catalogs	T-5830
	Faceted class	Faceted classification	T-5955
	I&R	Information and referral services	T-5161
	International Standard Book Number	ISBN -	T-5750
	LC Subject heading	LIBRARY OF CONGRESS SUBJECT HEADINGS	T-5655
	Robert Burton House Library	Robert Burton House Undergraduate Library	T-1240
	S.C. SEARS	South Carolina SEARS LIST OF SUB- JECT HEADINGS	T-7370 T-5660
	UNC-Chapel Hill	University of North Carolina at Chapel Hill	T-8832
	Video tapes	Videotapes	T-2835
	Syntactical variations		
Syntactical variations	School of Library Science	Library science schools	T-431
	Systems of security	Security systems	T-4531
	Training of staff	Staff training	T-4070
Semantically close	Cataloging convention	Cataloging rules	T-5710
	Census report	Bureau of the Census publications	T-3170
	Evaluation of personnel	Staff evaluation	T-3990
	Evaluating personnel	Staff evaluation	T-3990
	International documents	International publications	T-3240
	Library building	Library construction	T-4480
	Mechanical abstracting	Automated abstracting	T-5535
	Personnel evaluation	Staff evaluation	T-3990

TABLE 19.--Continued

1 Relation	2 Synonym	3 Descriptor	4 Concept class
NARROWER TERMS	Preschooler	Preschool children	T-1860
	UN document	United Nations publications	T-1860
	United States government publications	Federal publications	T-3110
	Want lists	Desiderata	T-6210
	Inclusion		
	Diaries	Manuscripts	T-3035
	Library boards	Governing boards	T-3690
	North American Indians	Indians	T-1885
	Novels	Fiction	T-1665
	Tours of the library	Library orientation	T-5111
BROADER TERMS	Two-year colleges	Junior colleges and technical schools	T-8700
	University libraries	Academic libraries	T-1210
	Person-activity		
	Bookdealer	Book selling	T-2425
	Supervisor	Supervision	T-3980
	Components of compound descriptors		
	Approval plan	Approval and gathering plans	T-6180
	Book marketing	Book marketing and promotion	T-2350
	Map	Maps and atlases	T-3045
	Preschool	Preschool and day care	T-8510
BROADER TERMS	Inclusion		
	Children's library	Elementary school libraries	T-1330
	Employment	Employment practices	T-3890
	Illumination	Book illumination	T-2150
	Information storage and retrieval	Information retrieval	T-5181
Activity-person	Administering	Library director	T-4260

TABLE 19.--Continued

1 Relation	2 Synonym	3 Descriptor	4 Concept class
RELATE TERMS			
Grammatical variations	Administer	Administration	T-3510
	Biographical	Biography	T-1684
	Educational	Education	T-1710
	Mathematical	Mathematics	T-3510
Semantic overlap	Centralized cataloging	Cooperative cataloging	T-5765
	Hearing impaired	Deaf	T-1960
	Longevity	Length of service	T-4020
	Theses	Dissertations	T-2740
	Training management	Leadership training	T-4110
	Training techniques	Staff training	T-4070

REFERENCES

- Aitchison, J., and Gilchrist, A. 1972. Thesaurus construction: A practical manual. London, England: Aslib.
- Bhattacharyya, K. 1974. The effectiveness of natural language in science indexing and retrieval. Journal of Documentation 30:235-254.
- Bobrow, D. G. 1967. Syntactic theory in computer implementations. In Automated language processing, ed. H. Borko, pp. 215-251. Information Science Series. New York: Wiley.
- Caplan, C.; Holley, B.; McDuffee, D.; Pearse, L.; and Stephenson, C. 1978. "Master's papers thesaurus: Thes 78." Typewritten. Chapel Hill, N.C.: University of North Carolina at Chapel Hill.
- Climenson, W. D.; Hardwick, N. H.; and Jacobson, S. N. 1961. Automatic syntax analysis in machine indexing and abstracting. In Machine indexing: progress and problems, pp. 305-325. Washington, D.C.: Center for Technology and Administration, American University.
- Committee on Scientific and Technical Information. 1967. Guidelines for the development of information retrieval thesauri. Washington, D.C.: Government Printing Office.
- Fishman, M. 1972. The transformational model of language and information retrieval. Drexel Library Quarterly 8:193-200.
- Hines, T. 1968. Vocabulary control in indexing the literature of librarianship and information science. Bethesda, Md.: ERIC Document Reproduction Service, ED 050 742.
- Hutchins, W. D. 1975. Languages of indexing and classification: A linguistic study of structures and functions. Librarianship and Information Studies, vol. 3. Stevenage, England: P. Peregrinus.
- Klingbiel, P. H. 1975. Machine-aided indexing of technical literature. Information Storage and Retrieval 9:79-84.
- Richmond, P. A. 1973. A thesaurus within a thesaurus: A study in ambiguity. In Toward a theory of librarianship: Papers in honor of Jesse Hauk Shera, ed. C. H. Rawski, pp. 268-301. Metuchen, N. J.: Scarecrow Press.

- Rolling, L. N. 1970. Compilation of thesauri for use in computer systems. Information Storage and Retrieval 6:341-350.
- Salton, G. 1968. Automatic information organization and retrieval. Computer Science Series. New York: McGraw-Hill.
- Salton, G. 1975. A theory of indexing. Regional Conference Series in Applied Mathematics, no. 15. Philadelphia: Society for Industrial and Applied Mathematics.
- Silver, I. D. P. 1974. "Thesaurus maintenance manual." Photoreproduced. Chapel Hill, N. C.: Carolina Population Center.
- Soergel, D. 1974. Indexing languages and thesauri: Construction and maintenance. Information Sciences Series. Los Angeles: Melville Pub. Co.
- Stevens, M. E. 1965. Automatic indexing: A state-of-the-art report. National Bureau of Standards Monograph 91. Washington, D.C.: Government Printing Office.
- Stone, P. J. 1969. Improved quality of content-analysis categories: Computerized disambiguation rules for high frequency English words. In The analysis of communication content: Developments in scientific theories and computer techniques, ed. G. Gerbner, pp. 199-221. New York: Wiley.
- Strode, P. S. 1976. "User's manual for the Automatic Indexing system." Master's thesis, University of North Carolina at Chapel Hill.
- Swanson, D. R. 1961. The nature of multiple meaning. In Proceedings of the national symposium on machine translation, ed. H. P. Edmundson, pp. 386-393. Englewood Cliffs, N. J.: Prentice-Hall.
- Wall, N. 1972. Thesauri and classifications in non-scientific areas. Drexel Library Quarterly 8:129-234.
- Weiss, S. F. 1973. Learning to disambiguate. Information Storage and Retrieval 9:33-41.